Behavioral and Brain Sciences

cambridge.org/bbs

The generalizability crisis

Tal Yarkoni

Department of Psychology, The University of Texas at Austin, Austin, TX 78712-1043, USA
tyarkoni@gmail.com

Target Article

Cite this article: Yarkoni T. (2022) The generalizability crisis. *Behavioral and Brain Sciences* **45**, e1: 1–78. doi:10.1017/S0140525X20001685

Target Article Accepted: 11 December 2020
Target Article Manuscript Online: 21 December 2020
Commentaries Accepted: 31 March 2021

Key words:
Generalization; inference; philosophy of science; psychology; random effects; statistics

Abstract

Most theories and hypotheses in psychology are verbal in nature, yet their evaluation overwhelmingly relies on inferential statistical procedures. The validity of the move from qualitative to quantitative analysis depends on the verbal and statistical expressions of a hypothesis being closely aligned – that is, that the two must refer to roughly the same set of hypothetical observations. Here, I argue that many applications of statistical inference in psychology fail to meet this basic condition. Focusing on the most widely used class of model in psychology – the linear mixed model – I explore the consequences of failing to statistically operationalize verbal hypotheses in a way that respects researchers' actual generalization intentions. I demonstrate that although the "random effect" formalism is used pervasively in psychology to model intersubject variability, few researchers accord the same treatment to other variables they clearly intend to generalize over (e.g., stimuli, tasks, or research sites). The under-specification of random effects imposes far stronger constraints on the generalizability of results than most researchers appreciate. Ignoring these constraints can dramatically inflate false-positive rates, and often leads researchers to draw sweeping verbal generalizations that lack a meaningful connection to the statistical quantities they are putatively based on. I argue that failure to take the alignment between verbal and statistical expressions seriously lies at the heart of many of psychology's ongoing problems (e.g., the replication crisis), and conclude with a discussion of several potential avenues for improvement.

1. Introduction

Modern psychology is – at least to superficial appearances – a quantitative discipline. Evaluation of most claims proceeds by computing statistical quantities that are thought to bear some important relationship to the theories or practical applications psychologists care about. This observation may seem obvious, but it's worth noting that things didn't have to turn out this way. Given that the theories and constructs psychologists are interested in usually have qualitative origins, and are almost invariably expressed verbally, a naive observer might well wonder why psychologists bother with numbers at all. Why take the trouble to compute *p*-values, Bayes factors, or confidence intervals when evaluating qualitative theoretical claims? Why don't psychologists simply look at the world around them, think deeply for a while, and then state – again in qualitative terms – what they think they have learned?

The standard answer to this question is that quantitative analysis offers important benefits that qualitative analysis cannot (e.g., Steckler, McLeroy, Goodman, Bird, & McCormick, 1992) – perhaps most notably, greater objectivity and precision. Two observers can disagree over whether a crowd of people should be considered "big" or "small," but if a careful count establishes that the crowd contains exactly 74 people, then it is at least clear what the facts on the ground are, and any remaining dispute is rendered largely terminological.

Unfortunately, the benefits of quantitation come at a steep cost: Verbally expressed psychological constructs[1] – things like cognitive dissonance, language acquisition, and working memory capacity – cannot be directly measured with an acceptable level of objectivity and precision. What *can* be measured objectively and precisely are operationalizations of those constructs – for example, a performance score on a particular digit span task, or the number of English words an infant has learned by age 3. Trading vague verbal assertions for concrete measures and manipulations is what enables researchers to draw precise, objective, quantitative inferences; however, the same move also introduces new points of potential failure, because the validity of the original verbal assertion now depends not only on what happens to be true about the world itself, but also on the degree to which the chosen proxy measures successfully capture the constructs of interest – what psychometricians term *construct validity* (Cronbach & Meehl, 1955; Guion, 1980; O'Leary-Kelly & Vokurka, 1998).

When the construct validity of a measure or manipulation is low, any conclusions one draws at the operational level run a high risk of failing to generalize to the construct level. An easy way to appreciate this is to consider an extreme example. Suppose I hypothesize that high social status makes people behave dishonestly. If I claim that I can test this hypothesis by randomly assigning people to either read a book or watch television for 10 min, and

CAMBRIDGE UNIVERSITY PRESS

then measuring their performance on a speeded dishwashing task, nobody is going to take me very seriously. It doesn't even matter how the results of my experiment turn out: There is no arrangement of numbers in a table, no *p*-value I could compute from my data, that could possibly turn my chosen experimental manipulation into a sensible proxy for social status. And the same goes for the rather questionable use of speeded dishwashing performance as a proxy for dishonesty.

The absurdity of the preceding example exposes a critical assumption that often goes unnoticed: For an empirical result to have bearing on a verbal assertion, the measured variables must be suitable operationalizations of the verbal constructs of interest, and the relationships between the measured variables must parallel those implied by the logical structure of the verbal statements. Equating the broad construct of honesty with a measure of speeded dishwashing is so obviously nonsensical that we immediately reject such a move out of hand. What may be less obvious is that exactly the same logic implicitly applies in virtually every case where researchers lean on statistical quantities to justify their verbal claims. Statistics is not, as many psychologists appear to view it, a rote, mechanical procedure for turning data into conclusions. It is better understood as a parallel, and more precise, *language* in which one can express one's hypotheses or beliefs. Every statistical model is a description of some real or hypothetical state of affairs in the world. If its mathematical expression fails to capture roughly the same state of affairs as the verbal hypothesis the researcher began with, then the statistical quantities produced by the model cannot serve as an adequate proxy for the verbal statements – and consequently, the former cannot be taken as support for the latter.

Viewed from this perspective, the key question is how closely the verbal and quantitative expressions of one's hypothesis align with each other. When a researcher verbally expresses a particular proposition – be it a theoretically informed hypothesis or a purely descriptive characterization of some data – she is implicitly defining a set of hypothetical measurements (or *admissible observations*; Brennan, 1992) that would have to come out a certain way in order for the statement to be corroborated. If the researcher subsequently asserts that a particular statistical procedure provides a suitable operationalization of the verbal statement, she is making the tacit but critical assumption that the universe of hypothetical measurements implicitly defined by the chosen statistical procedure, in concert with the experimental design and measurement model, is well aligned with the one implicitly defined by the qualitative statement. Should a discrepancy between the two be discovered, the researcher will then face a choice between (a) working to resolve the discrepancy in some way (i.e., by modifying either the verbal statement or the quantitative procedure(s) meant to provide an operational parallel); or (b) giving up on the link between the two and accepting that the statistical procedure does not inform the verbal expression in a meaningful way.

The next few sections explore this relationship with respect to the most widely used class of statistical model in psychology – linear mixed models containing fixed and random effects (although the broader conceptual points I will make apply to *any* use of statistical quantities to evaluate verbal claims). The exploration begins with an examination of the standard random-subjects model – a mainstay of group-level inferences in psychology – and then progressively considers additional sources of variability whose existence is implied by most verbal inferences in psychology, but that the standard model fails to appropriately capture. The revealed picture is that an unknown but clearly very large fraction of statistical hypotheses described in psychology studies cannot plausibly be considered reasonable operationalizations of the verbal hypotheses they are meant to inform. (Although I deliberately restrict the focus of my discussion to the field of psychology, with which I am most familiar, I expect that researchers in various social and biomedical disciplines will find that the core arguments I lay out generalize well to many other areas.)

Tal Yarkoni is a Research Associate Professor at the University of Texas at Austin. He writes journal articles, software, and short fiction, and enjoys eating ice cream and being a general nuisance.

2. Fixed versus random effects

Let us begin with a scenario that will be familiar to many psychologists. Suppose we administer a cognitive task – say, the color-word Stroop (MacLeod, 1991; Stroop, 1935) – to a group of participants (the reader is free to mentally substitute almost any other experimental psychology task into the example). Each participant is presented with a series of trials, half in a congruent condition and half in an incongruent condition. We are tasked with fitting a statistical model to estimate the canonical Stroop effect – that is, the increase in reaction time (RT) observed when participants are presented with incongruent color-word information relative to congruent color-word information.

A naive, although almost always inappropriate, model might be the following:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 X_{ij} + e_{ij} \\ e_{ij} &\sim \mathcal{N}(0, \sigma_e^2) \end{aligned} \tag{1}$$

In this linear regression, y_{ij} denotes the $_i$-th subject's response on trial $_j$, X_{ij} indexes the experimental condition (congruent or incongruent) of subject i's j-th trial, β_0 is an intercept, β_1 is the effect of congruency, and e_{ij} captures the errors, which are assumed to be normally distributed.

What is wrong with this model? Well, one rather serious problem is that the model blatantly ignores sources of variance in the data that we know on theoretical grounds must exist. Notably, because the model includes only a single intercept parameter and a single slope parameter across all subjects and trials, it predicts exactly the same RT value for all trials in each condition, no matter which subject a given trial is drawn from. Such an assumption is clearly untenable: It's absurd to suppose that the only source of trial-to-trial RT variability within experimental conditions is random error. We know full well that people differ systematically from one another in performance on the Stroop task (and for that matter, on virtually every other cognitive task). Any model that fails to acknowledge this important source of variability is clearly omitting an important feature of the world as we understand it.

From a statistical standpoint, the model's failure to explicitly acknowledge between-subject variability has several deleterious consequences for our Stroop estimate. The most salient one, given psychologists' predilection toward dichotomous conclusions (e.g., whether or not an effect is statistically significant), is that the estimated uncertainty surrounding the parameter estimates of interest will tend to be biased – typically downward (i.e., in our Stroop example, the standard error of the Stroop effect will usually be underestimated).[2] The reason is that, lacking any concept of a

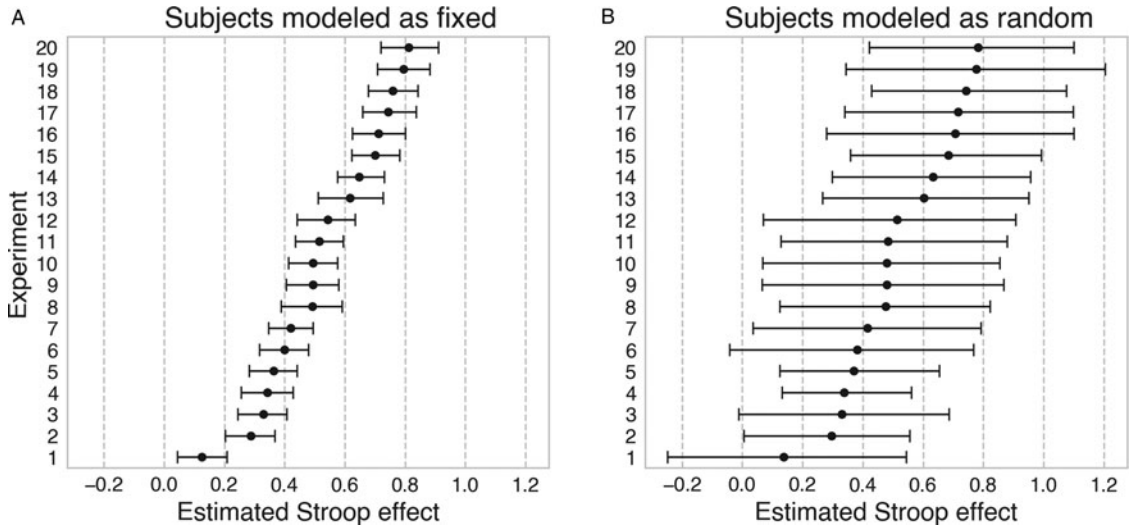

Figure 1. Consequences of mismatch between model specification and generalization intention. Each row represents a simulated Stroop experiment with $n = 20$ new subjects randomly drawn from the same global population (the ground truth for all parameters is constant over all experiments). Bars display the estimated Bayesian 95% highest posterior density (HPD) intervals for the (fixed) condition effect of interest in each experiment. Experiments are ordered by the magnitude of the point estimate for visual clarity. (A) The fixed-effects model specification in Eq. (1) does not account for random subject sampling, and consequently underestimates the uncertainty associated with the effect of interest. (B) The random-effects specification in Eq. (2) takes subject sampling into account, and produces appropriately calibrated uncertainty estimates.

"person," our model cannot help but assume that any new set of trials – no matter who they come from – must have been generated by exactly the same set of processes that gave rise to the trials the model has previously seen. Consequently, the model cannot adjust the uncertainty around the point estimate to account for variability between subjects, and will usually produce an overly optimistic estimate of its own performance when applied to new subjects whose data-generating process is at least somewhat different from the process that generated the data the model was trained on.

The deleterious impact of using model (1) to estimate the Stroop effect when generalization to new subjects is intended is illustrated in Figure 1A. The figure shows the results of a simulation of 20 random Stroop experiments, each with 20 participants and 200 trials per participant (100 in each condition). The true population effect – common to all 20 experiments – is assumed to be large. As expected, fitting the simulated data with the fixed-effects model specification in Eq. (1) produces an unreasonably narrow estimate of the uncertainty surrounding the point estimates – observe that, for any given experiment, most of the estimates from the other experiments are well outside the 95% highest posterior density (HPD) interval. Researchers who attempt to naively generalize the estimates obtained using the fixed-effects model to data from new subjects are thus setting themselves up for an unpleasant surprise.

How might we adjust model (1) to account for the additional between-subject variance in the data introduced by the stochastic sampling of individuals from a broader population? One standard approach is to fit a model as shown below:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 X_{ij} + u_{0i} + u_{1i} X_{ij} + e_{ij} \\ u_{0i} &\sim \mathcal{N}(0, \sigma^2_{u_0}) \\ u_{1i} &\sim \mathcal{N}(0, \sigma^2_{u_1}) \\ e_{ij} &\sim \mathcal{N}(0, \sigma^2_{e}) \end{aligned} \tag{2}$$

Here, we expand model (1) to include two new terms: u_0 and u_1, which, respectively, reflect a set of intercepts and a set of slopes – one pair of terms per subject.[3] The u parameters are assumed (like the error e) to follow a normal distribution centered at zero, with the size of the variance components (i.e., the variances of the groups of random effects) $\sigma^2_{u_k}$ estimated from the data.

Conventionally, the u parameters in model (2) are referred to as *random* (or sometimes, *varying* or *stochastic*) effects, as distinct from the *fixed* effects captured by the β terms.[4] There are several ways to conceptualize the distinction between random and fixed effects (Gelman & Hill, 2006), but because of our focus here is on generalizability, we will define them this way: Fixed effects are used to model variables that must remain constant in order for the model to preserve its meaning across replication studies; random effects are used to model indicator variables that are assumed to be stochastically sampled from some underlying population and can vary across replications without meaningfully altering the research question. In the context of our Stroop example, we can say that the estimated Stroop effect β_1 is a fixed effect, because if we were to run another experiment using a different manipulation (say, a Sternberg memory task), we could no longer reasonably speak of the second experiment being a replication of the first. By contrast, psychologists almost invariably think of experimental subjects as a random factor: we are rarely interested in the particular people we happen to have in a given sample, and it would be deeply problematic if two Stroop experiments that differed only in their use of different subjects (randomly sampled from the same population) had to be treated as if they provided estimates of two conceptually distinct Stroop effects.[5]

Note that although the model specified in (2) is a substantial improvement over the one specified in (1) if our goal is to draw inferences over populations of subjects, it is not in any meaningful sense the "correct" model. Model (2) is clearly still an extremely simplistic approximation of the true generative processes underlying Stroop data, and, even within the confines of purely linear models, there are many ways in which we could further elaborate

on (2) to account for other potentially important sources of variance (e.g., practice or fatigue effects, stimulus-specific effects, measured individual differences in cognitive ability, etc.). Moreover, the fact that model (2) supports inference over *some* broader population of subjects provides no guarantee that that population is one the researcher is interested in. If, for example, our subjects are all sampled from a Western undergraduate population aged 18–23, then model (2) may license generalization of the results to other undergraduates like the ones we studied, but we would be leaning very heavily on auxiliary assumptions not explicitly included in our model if we were to generalize our conclusions to the broader population of human beings.

In highlighting the difference between models (1) and (2), I simply wish to draw attention to two important and interrelated points. First, inferences about model parameters are always tied to a particular model specification. A claim like "there is a statistically significant effect of Stroop condition" is not a claim about the world *per se*; rather, it is a claim about the degree to which a specific model accurately describes the world under certain theoretical assumptions and measurement conditions. Strictly speaking, a statistically significant effect of Stroop condition in model (1) tells us only that the data we observe would be unlikely to occur under a null model that considers all trials to be completely exchangeable. By contrast, a statistically significant effect in model (2) for what nominally appears to be the "same" β_1 parameter would have a different (and somewhat stronger) interpretation, as we are now entitled to conclude that the data we observe would be unlikely if there were no effect (on average) at the level of individuals randomly drawn from some population.

Second, the validity of an inference depends not just on the model itself, but also on the analyst's (typically implicit) intentions. As discussed earlier, to support valid inference, a statistical model must adequately represent the universe of observations the analyst intends to implicitly generalize over when drawing qualitative conclusions. In our example above, what makes model (1) a bad model is not the model specification alone, but the fact that the specification aligns poorly with the universe of observations that researchers typically care about. In typical practice, researchers intend their conclusions to apply to entire populations of subjects, and not just to the specific individuals who happened to walk through the laboratory door when the study was run. Critically, then, *it is the mismatch between our generalization intention and the model specification that introduces an inflated risk of inferential error, and not the model specification alone.* The reason we model subjects as random effects is not that such a practice is objectively better, but rather, that this specification more closely aligns the meaning of the quantitative inference with the meaning of the qualitative hypothesis we're interested in evaluating (for discussion, see Cornfield & Tukey, 1956).

3. Beyond random subjects

The discussion in the preceding section may seem superfluous to some readers given that, in practice, psychologists almost universally already model subject as a random factor in their analyses. Importantly, however, there is nothing special about subjects. In principle, what goes for subjects also holds for any other factor of an experimental or observational study whose levels the authors intend to generalize over. The reason that we routinely inject extra uncertainty into our models in order to account for between-subject variability is that we want our conclusions to apply to a broader population of individuals, and not just to the specific people we randomly sampled. But the same logic also applies to a large number of other factors that we do not routinely model as random effects – stimuli, experimenters, research sites, and so on. Indeed, as Brunswik long ago observed, "proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another" (Brunswik, 1947, p. 179). As we shall see, extending the random effects treatment to other factors besides subjects has momentous implications for the interpretation of a vast array of published findings in psychology.

3.1. The stimulus-as-fixed-effect fallacy

A paradigmatic example of a design factor that psychologists almost universally – and inappropriately – model as a fixed rather than random factor is experimental stimuli. The tendency to ignore stimulus sampling variability has been discussed in the literature for over 50 years (Baayen, Davidson, & Bates, 2008; Clark, 1973; Coleman, 1964; Judd, Westfall, & Kenny, 2012), and was influentially dubbed the *fixed-effect fallacy* by (Clark, 1973). Unfortunately, outside of a few domains such as psycholinguistics, it remains rare to see psychologists model stimuli as random effects – despite the fact that most inferences researchers draw are clearly meant to generalize over populations of stimuli. The net result is that, strictly speaking, the inferences routinely drawn throughout much of psychology can only be said to apply to a specific – and usually small – set of stimuli. Generalization to the broader class of stimuli like the ones used is not licensed.

It is difficult to overstate how detrimental an impact the stimulus-as-fixed-effect fallacy has had – and continues to have – in psychology. Empirical studies in domains ranging from social psychology to functional magnetic resonance imaging (MRI) have demonstrated that test statistic inflation of up to 300% is not uncommon, and that, under realistic assumptions, false-positive rates in many studies could easily exceed 60% (Judd et al., 2012; Westfall, Nichols, & Yarkoni, 2016; Wolsiefer, Westfall, & Judd, 2017). In cases where subject sample sizes are very large, stimulus samples are very small, and stimulus variance is large, the false-positive rate theoretically approaches 100%.

The clear implication of such findings is that many literatures within psychology are likely to be populated by studies that have spuriously misattributed statistically significant effects to fixed effects of interest when they should actually be attributed to stochastic variation in uninteresting stimulus properties. Moreover, given that different sets of stimuli are liable to produce effects in opposite directions (e.g., when randomly sampling 20 nouns and 20 verbs, some samples will show a statistically significant noun > verb effect, whereas others will show the converse), it is not hard to see how one could easily end up with entire literatures full of "mixed results" that seem statistically robust in individual studies, yet cannot be consistently replicated across studies.

3.2. Generalizing the generalizability problem

The stimulus-as-fixed-effect fallacy is but one special case of a general trade-off between precision of estimation and breadth of generalization. Each additional random factor one adds to a model licenses generalization over a corresponding population of potential measurements, expanding the scope of inference beyond only those measurements that were actually obtained.

However, adding random factors to one's model also typically increases the uncertainty with which the fixed effects of interest are estimated. The fact that most psychologists have traditionally modeled only subject as a random factor – and have largely ignored the variance introduced by stimulus sampling – is probably best understood as an accident of history (or, more charitably perhaps, of technological limitations, as the software and computing resources required to fit such models were hard to come by until fairly recently).

Unfortunately, just as the generalizability problem doesn't begin and end with subjects, it also doesn't end with subjects and stimuli. Exactly the same considerations apply to all other aspects of one's experimental design or procedure that could, in principle, be varied without substantively changing the research question. Common design factors that researchers hardly ever vary, yet almost invariably intend to generalize over, include experimental task, between-subject instructional manipulation, research site, experimenter (or, in clinical studies, therapist; e.g., Crits-Christoph & Mintz, 1991), instructions, laboratory testing conditions (e.g., Crabbe, Wahlsten, & Dudek, 1999; Wahlsten et al., 2003), weather, and so on and so forth effectively *ad infinitum*.

Naturally, the degree to which each such factor matters will vary widely across domain and research question. I'm not suggesting that most statistical inferences in psychology are invalidated by researchers' failure to explicitly model what their participants ate for breakfast 3 days prior to participating in a study. Collectively, however, unmodeled factors almost always contribute substantial variance to the outcome variable. Failing to model such factors appropriately (or at all) means that a researcher will end up either (a) running studies with substantially higher-than-nominal false-positive rates, or (b) drawing inferences that technically apply only to very narrow, and usually uninteresting, slices of the universe the researcher claims to be interested in.

3.3. Case study: verbal overshadowing

To illustrate the problem, it may help to consider an example. Alogna and colleagues (2014) conducted a large-scale "registered replication report" (RRR; Simons, Holcombe, & Spellman, 2014) involving 31 sites and over 2,000 participants. The study sought to replicate an influential experiment by Schooler and Engstler-Schooler (1990) in which the original authors showed that participants who were asked to verbally describe the appearance of a perpetrator caught committing a crime on video showed poorer recognition of the perpetrator following a delay than did participants assigned to a control task (naming as many countries and capitals as they could). Schooler and Engstler-Schooler (1990) dubbed this the *verbal overshadowing effect*. In both the original and replication experiments, only a single video, containing a single perpetrator, was presented at encoding, and only a single set of foil items was used at test. Alogna et al. successfully replicated the original result in one of two tested conditions, and concluded that their findings revealed "a robust verbal overshadowing effect" in that condition.

Let us assume for the sake of argument that there is a genuine and robust causal relationship between the manipulation and outcome employed in the Alogna et al. study. I submit that there would still be essentially no support for the authors' assertion that they found a "robust" verbal overshadowing effect, because the experimental design and statistical model used in the study simply cannot support such a generalization. The strict conclusion we are entitled to draw, given the limitations of the experimental design inherited from Schooler and Engstler-Schooler (1990), is that there is at least one particular video containing one particular face that, when followed by one particular lineup of faces, is more difficult for participants to identify if they previously verbally described the appearance of the target face than if they were asked to name countries and capitals. This narrow conclusion does not preclude the possibility that the observed effect is specific to this one particular stimulus, and that many other potential stimuli the authors could have used would have eliminated or even reversed the observed effect. (In later sections, I demonstrate that the latter conclusion is statistically bound to be true given even very conservative background assumptions about the operationalization, and also that one can argue from first principles – i.e., *without any data at all* – that there must be *many* stimuli that show a so-called verbal overshadowing effect.)

Of course, stimulus sampling is not the only unmodeled source of variability we need to worry about. We also need to consider any number of other plausible sources of variability: research site, task operationalization (e.g., timing parameters, modality of stimuli or responses), instructions, and so on. On any reasonable interpretation of the *construct* of verbal overshadowing, the corresponding universe of intended generalization should clearly also include most of the operationalizations that would result from randomly sampling various combinations of these factors (e.g., one would expect it to still count as verbal overshadowing if Alogna et al. had used live actors to enact the crime scene, instead of showing a video).[6] Once we accept this assumption, however, the critical question researchers should immediately ask themselves is: Are there other psychological processes besides verbal overshadowing that could plausibly be influenced by random variation in any of these uninteresting factors, *independently of the hypothesized psychological processes of interest*? A moment or two of consideration should suffice to convince one that the answer is a resounding yes. It is not hard to think of dozens of explanations unrelated to verbal overshadowing that could explain the causal effect of a given manipulation on a given outcome in any single operationalization.[7]

This verbal overshadowing example is by no means unusual. The same concerns apply equally to the broader psychology literature containing tens or hundreds of thousands of studies that routinely adopt similar practices. In most of psychology, it is standard operating procedure for researchers employing just one experimental task, between-subject manipulation, experimenters, testing room, research site, and so on, to behave as though an extremely narrow operationalization is an acceptable proxy for a much broader universe of admissible observations. It is instructive – and somewhat fascinating from a sociological perspective – to observe that although no psychometrician worth their salt would ever recommend a default strategy of measuring complex psychological constructs using a single unvalidated item, the majority of psychology studies do precisely that with respect to *multiple* key design factors. The modal approach is to stop at a perfunctory demonstration of face validity – that is, to conclude that if a particular operationalization *seems* like it has something to do with the construct of interest, then it is an acceptable stand-in for that construct. Any measurement-level findings are then uncritically generalized to the construct level, leading researchers to conclude that they've learned something useful about broader phenomena like verbal overshadowing, working memory, ego depletion, and so on, when in fact such sweeping generalizations typically obtain little support from the reported empirical studies.

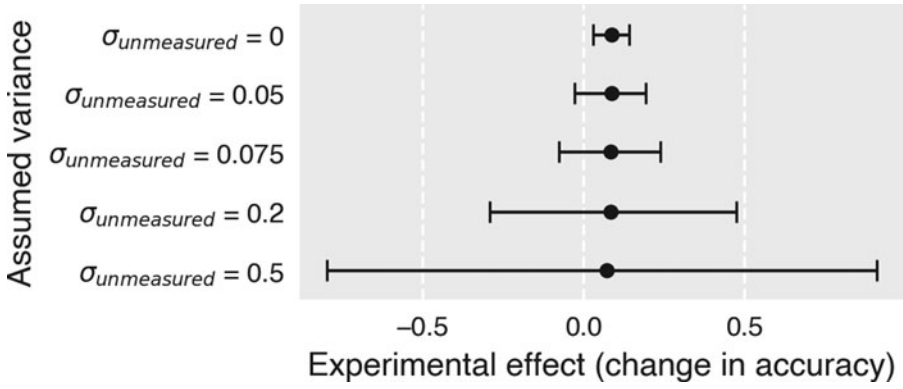

Figure 2. Effects of unmeasured variance components on the putative "verbal overshadowing" effect. Error bars display the estimated Bayesian 95% highest posterior density (HPD) intervals for the experimental effect reported in Alogna et al. (2014). Positive estimates indicate better performance in the control condition than in the experimental condition. Each row represents the estimate from the model specified in Eq. (4), with only the size of $\sigma^2_{\text{unmeasured}}$ (corresponding to $\sigma^2_{u_2}$ in Eq. (4)) varying as indicated. This parameter represents the assumed contribution of all variance components that are unmeasured in the experiment, but fall within the universe of intended generalization conceptually. The top row ($\sigma^2_{u_2} = 0$) can be interpreted as a conventional model analogous to the one reported in Alogna et al. (2014) – that is, it assumes that no unmeasured sources have any impact on the putative verbal overshadowing effect.

4. Unmeasured factors

In an ideal world, generalization failures like those described above could be addressed primarily via statistical procedures – for example, by adding new random effects to models. In the real world, this strategy is a non-starter: In most studies, the vast majority of factors that researchers intend to implicitly generalize over don't actually observably vary in the data, and therefore can't be accounted for using traditional mixed-effects models. Unfortunately, the fact that one has failed to introduce or measure variation in one or more factors doesn't mean those factors can be safely ignored. Any time one samples design elements into one's study from a broader population of possible candidates, one introduces sampling error that is likely to influence the outcome of the study to some unknown degree.

Suppose we generalize our earlier model (2) to include all kinds of random design factors that we have no way of directly measuring:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 X_{1ij} + u_{0ij} + u_{1ij} + \cdots + u_{kij} + e_{ij} \\ u_{kij} &\sim \mathcal{N}(0, \sigma^2_{u_k}) \\ e_{ij} &\sim \mathcal{N}(0, \sigma^2_e) \end{aligned} \tag{3}$$

Here, $u_0 \ldots u_k$ are placeholders for all of the variance components that we implicitly consider part of the universe of admissible observations, but that we have no way of measuring or estimating in our study. It should be apparent that our earlier model (2) is just a special case of (3) where the vast majority of the u_k and $\sigma^2_{u_k}$ terms are fixed to 0. That is – and this is arguably the most important point in this paper – the conventional "random-effects" model (where in actuality only subjects are modeled as random effects) *assumes exactly zero effect of site, experimenter, stimuli, task, instructions, and every other factor except subject* – even though in most cases it's safe to assume that such effects exist and are non-trivial, and even though authors almost invariably start behaving as if their statistical models did, in fact, account for such effects as soon as they reach the "Discussion" section.

4.1. Estimating the impact

We do not have to take the urgency of the above exhortation on faith. Although it's true that we can't directly estimate the population magnitude of variance components that showed no observable variation in our sample, we can still simulate their effects under different assumptions. Doing so allows us to demonstrate empirically that even modest assumptions about the magnitude of unmeasured variance components may be sufficient to completely undermine many conventional inferences about fixed effects of interest.

To illustrate, let's return to Alogna et al.'s (2014) verbal overshadowing RRR.

Recall that the dataset included data from over 2,000 subjects sampled at 31 different sites, but used exactly the same experimental protocol (including the same single stimulus sequence) at all sites. Because of most of the data are publicly available, we can fit a mixed-effects model to try and replicate the reported finding of a "robust verbal overshadowing effect." Both the dataset and the statistical model used here differ somewhat from the ones in Alogna et al. (2014),[8] but the differences are immaterial for our purposes. As Figure 2 illustrates (top row, labeled $\sigma^2_{\text{unmeasured}} = 0$), we can readily replicate the key finding from Alogna et al. (2014): Participants assigned to the experimental condition were more likely to misidentify the perpetrator seen in the original video.

We now ask the following question: How would the key result depicted in the top row of Figure 2 changes if we *knew* the size of the variance component associated with random stimulus sampling? This question cannot be readily answered using classical inferential procedures (because there's only a single stimulus in the dataset, so the variance component is non-identifiable), but is trivial to address using a Bayesian estimation framework. Specifically, we fit the following model:

$$\begin{aligned} y_{ps} &= \beta_0 + \beta_1 X_{ps} + u_{0s} + u_{1s} X_{ps} + u_2 X_{ps} + e_{ps} \\ u_{0s} &\sim \mathcal{N}(0, \sigma^2_{u_0}) \\ u_{1s} &\sim \mathcal{N}(0, \sigma^2_{u_1}) \\ u_2 &\sim \mathcal{N}(0, \sigma^2_{u_2}) \\ e_{ps} &\sim \mathcal{N}(0, \sigma^2_e) \end{aligned} \tag{4}$$

Here, p indexes participants, s indexes sites, X_{ps} indexes the experimental condition assigned to participant p at site s, the β terms encode the fixed intercept and condition slope, and the u terms encode the random effects (site-specific intercepts u_0, site-specific slopes u_1, and the stimulus effect u_2). The novel feature of this model is the inclusion of u_2, which would ordinarily reflect the variance in outcome associated with random stimulus sampling, but is constant in our dataset (because there's only a single stimulus). Unlike the other parameters, we cannot estimate u_2 from the data. Instead, we fix its prior during estimation, by setting

$\sigma^2_{u_2}$ to a specific value. Although the posterior estimate of $_{u2}$ is then necessarily identical to its prior (because the prior makes no contact with the data), and so is itself of no interest, the inclusion of the prior has the incidental effect of (appropriately) increasing the estimation uncertainty around the fixed effect(s) of interest. Conceptually, one can think of the added prior as a way of quantitatively representing our uncertainty about whether any experimental effect we observe should really be attributed to verbal overshadowing *per se*, as opposed to irrelevant properties of the specific stimulus we happened to randomly sample into our experiment. By varying the amount of variance injected in this way, we can study the conditions under which the conclusions obtained from the "standard" model (i.e., one that assumes zero effect of stimuli) would or wouldn't hold.

As it turns out, injecting even a small amount of stimulus sampling variance to the model has momentous downstream effects. If we very conservatively set $\sigma^2_{u_2}$ to 0.05, the resulting posterior distribution for the condition effect expands to include negative values within the 95% HPD (Fig. 2). For perspective, 0.05 is considerably lower than the between-site variance estimated from these data ($\sigma^2_{u_1} = 0.075$) – and it's quite unlikely that there would be less variation between different stimuli at a given site than between different sites for the same stimulus (as reviewed above, in most domains where stimulus effects have been quantitatively estimated, they tend to be large). Thus, even under very conservative assumptions about how much variance might be associated with stimulus sampling, there is little basis for concluding that there is a *general* verbal overshadowing effect. To draw Alogna et al.'s conclusion that there is a "robust" verbal overshadowing effect, one must effectively equate the construct of verbal overshadowing with *almost exactly* the operationalization tested by Alogna et al. (and Schooler & Schooler-Engstler before that), down to the same single video.

Of course, stimulus variance isn't the only missing variance component we ought to worry about. As Eq. (3) underscores, many other components are likely to contribute non-negligible variance to outcomes within our universe of intended generalization. We could attempt to list these components individually and rationally estimate their plausible magnitudes if we like, but an alternative route is to invent an omnibus parameter, $\sigma^2_{\text{unmeasured}}$, that subsumes *all* of the unmeasured variance components we expect to systematically influence the condition estimate β_1. Then we can repeat our estimation of the model in Eq. (4) with larger values of $\sigma^2_{u_2}$ (for the sake of convenience, I treat $\sigma^2_{u_2}$ and $\sigma^2_{\text{unmeasured}}$ interchangeably, as the difference is only that the latter is larger than the former).

For example, suppose we assume that the hypothetical aggregate influence of all the unmodeled variance components roughly equals the residual within-site variance estimated in our data (i.e., $\sigma^2_{\text{unmeasured}}$). This is arguably still fairly conservative when one considers that the aggregate $\sigma^2_{\text{unmeasured}}$ now includes not only stimulus sampling effects, but also the effects of differences in task operationalization, instructions, and so on. In effect, we are assuming that the net contribution of all of the uninteresting factors that vary across the entire universe of observations we consider "verbal overshadowing" is no bigger than the residual error we observe for this one particular operationalization. Yet fixing $\sigma^2_{\text{unmeasured}}$ to 0.5 renders our estimate of the experimental effect essentially worthless: the 95% HPD interval for the putative verbal overshadowing effect now spans values between −0.8 and 0.91 – almost the full range of possible values! The upshot is that, even given very conservative background assumptions, the massive Alogna et al. study – an initiative that drew on the efforts of dozens of researchers around the world – does not tell us much about the general phenomenon of verbal overshadowing. Under more realistic assumptions, it tells us essentially nothing. The best we can say, if we are feeling optimistic, is that it might tell us something about one particular *operationalization* of verbal overshadowing.[9]

The rather disturbing implication of all this is that, in any research area where one expects the aggregate contribution of the missing σ^2_u terms to be large – that is, anywhere that "contextual sensitivity" (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016) is high – the inferential statistics generated from models like (2) will often underestimate the true uncertainty surrounding the parameter estimates to such a degree as to make an outright mockery of the effort to learn something from the data using conventional inferential tests. Recall that the nominal reason we care about whether subjects are modeled as fixed or random effects is that the latter specification allows us to generalize to theoretically exchangeable observations (e.g., new subjects sampled from the same population), whereas the former does not. In practice, however, the majority of psychologists have no compunction about verbally generalizing their results not only to previously unseen subjects, but also to all kinds of other factors that have not explicitly been modeled – to new stimuli, experimenters, research sites, and so on.

Under such circumstances, it's unclear why anyone should really care about the inferential statistics psychologists report in most papers, seeing as those statistics bear only the most tenuous of connections to authors' sweeping verbal conclusions. Why take pains to ensure that subjects are modeled in a way that affords generalization beyond the observed sample – as nearly all psychologists reflexively do – whereas raising no objection whatsoever when researchers freely generalize their conclusions across all manner of variables that weren't explicitly included in the model at all? Why not simply model *all* experimental factors, including subjects, as fixed effects – a procedure that would, in most circumstances, substantially increase the probability of producing the sub-0.05 *p*-values psychologists so dearly crave? Given that we've already resolved to run roughshod over the relationship between our verbal theories and their corresponding quantitative specifications, why should it matter if we sacrifice the sole remaining sliver of generality afforded by our conventional "random-effects" models on the altar of the Biggest Possible Test Statistic?

It's hard to think of a better name for this kind of behavior than what Feynman famously dubbed *cargo cult science* (Feynman, 1974) – an obsessive concern with the superficial form of a scientific activity rather than its substantive empirical and logical content. Psychologists are trained to believe that their ability to draw meaningful inferences depends to a large extent on the production of certain statistical quantities (e.g., *p*-values below 0.05, Bayes Factor above 10, etc.), so they go to great effort to produce such quantities. That these highly contextualized numbers typically have little to do with the broad verbal theories and hypotheses that researchers hold in their heads, and take themselves to be testing, does not seem to trouble most researchers much. The important thing, it appears, is that the numbers have the right form.

5. A crisis of replicability or of generalizability?

It is worth situating the above concerns within the broader ongoing "replication crisis" in psychology and other sciences (Lilienfeld,

2017; Pashler & Wagenmakers, 2012; Shrout & Rodgers, 2018). My perspective on the replicability crisis broadly accords with other commentators who have argued that the crisis is real and serious, in the sense that there is irrefutable evidence that questionable research practices (Gelman & Loken, 2013; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011) and strong selection pressures (Francis, 2012; Kühberger, Fritz, & Scherndl, 2014; Smaldino & McElreath, 2016) have led to the publication of a large proportion of spurious or inflated findings that are unlikely to replicate (Ioannidis, 2005, 2008; Yarkoni, 2009). Accordingly, I think the ongoing shift toward practices such as preregistration, reporting checklists, data sharing, and so on, is a welcome development that will undoubtedly help improve the reproducibility and replicability of psychology findings.

At the same time, the current focus on reproducibility and replicability risks distracting us from more important, and logically antecedent, concerns about generalizability. The root problem is that when the manifestation of a phenomenon is highly variable across potential measurement contexts, it simply does not matter very much whether any single realization is replicable or not (cf. Gelman, 2015, 2018). Ongoing efforts to ensure the superficial reproducibility and replicability of effects – that is, the ability to obtain a similar-looking set of numbers from independent studies – are presently driving researchers in psychology and other fields to expend enormous resources on studies that are likely to have very little informational value even in cases where results can be consistently replicated. This is arguably clearest in the case of large-scale "registered replication reports" (RRRs) that have harnessed the enormous collective efforts of dozens of labs (e.g., Acosta et al., 2016; Alogna et al., 2014; Cheung et al., 2016; Eerland et al., 2016) – only to waste that collective energy on direct replications of a handful of poorly-validated experimental paradigms.

Although there is no denying that large, collaborative efforts could have enormous potential benefits (and there are currently a number of promising initiatives, for example, the Psychological Science Accelerator [Moshontz et al., 2018] and ManyBabies Consortium [Bergelson et al., 2017]), realizing these benefits will require a willingness to eschew direct replication in cases where the experimental design of the to-be-replicated study is fundamentally uninformative. Researchers must be willing to look critically at previous studies and flatly reject – on logical and statistical, rather than empirical, grounds – assertions that were never supported by the data in the first place, even under the most charitable methodological assumptions. A recognition memory task that uses just one video, one target face, and one set of foils simply cannot provide a meaningful test of a broad construct like verbal overshadowing, and it does a disservice to the field to direct considerable resources to the replication of such study. The appropriate response to a study like Schooler and Engstler-Schooler (1990) is to point out that the very narrow findings the authors reported did not – and indeed, could not, no matter how the data came out – actually support the authors' sweeping claims. Consequently, the study does not deserve any follow-up until such time as its authors can provide more compelling evidence that a phenomenon of any meaningful generality is being observed.

The same concern applies to many other active statistical and methodological debates. Is it better to use a frequentist or a Bayesian framework for hypothesis testing (Kruschke & Liddell, 2017; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007)? Should we move the conventional threshold for statistical significance from 0.05 to 0.005 (Benjamin et al., 2018; Lakens et al., 2018; McShane, Gal, Gelman, Robert, & Tackett, 2019)? A lot of ink continues to be spilled over such issues, yet in any research area where effects are highly variable (i.e., in most of psychology), the net contribution of such methodological and analytical choices to overall inferential uncertainty is likely to be dwarfed by the bias introduced by implicitly generalizing over unmodeled sources of variance in the data. There is little point in debating the merits of a statistical significance cut-off of 0.005 rather than 0.05 in a world where even a trivial change in an unmodeled variable – for example, a random choice between two nominally equivalent cognitive tasks, or the use of a slightly different stimulus sample – can routinely take one from $p = 0.5$ to $p = 0.0005$ or vice versa (cf. Crits-Christoph & Mintz, 1991; Westfall et al., 2016; Wolsiefer et al., 2017). Yet this root problem continues to go largely ignored in favor of efforts to treat its downstream symptoms. It appears that, faced with the difficulty of stating what the complex, multicausal effects we psychologists routinely deal in actually *mean*, we have collectively elected to instead pursue superficially precise answers to questions none of us really care much about.

To be clear, my suggestion is not that researchers should stop caring about methodological or statistical problems that presently limit reproducibility and replicability. Such considerations are undeniably important. My argument, rather, is that these considerations should be reserved for situations where the verbal conclusions drawn by researchers demonstrably bear some non-trivial connection to the reported quantitative analyses. The mere fact that a previous study has had a large influence on the literature is not a sufficient reason to expend additional resources on replication. On the contrary, the recent movement to replicate influential studies using more robust methods risks making the situation worse, because in cases where such efforts superficially "succeed" (in the sense that they obtain a statistical result congruent with the original), researchers then often draw the incorrect conclusion that the new data corroborate the original claim (e.g., Alogna et al., 2014) – when in fact the original claim was never supported by the data in the first place. A more appropriate course of action in cases where there are questions about the internal coherence and/or generalizability of a finding is to first focus a critical eye on the experimental design, measurement approach, and model specification. Only if a careful review suggests that these elements support the claims made by a study's authors should researchers begin to consider conducting a replication.

6. Where to from here?

A direct implication of the arguments laid out above is that a huge proportion of the quantitative inferences drawn in the published psychology literature is so weak as to be at best questionable and at worst utterly nonsensical. The difficult question I take up now is what we ought to do about this. I suggest three broad and largely disjoint courses of action researchers can pursue that would, in the aggregate, considerably improve the quality of research in psychological science.

6.1. Do something else

One perfectly reasonable course of action when faced with the difficulty of extracting meaningful, widely generalizable conclusions from effects that are inherently complex and highly variable is to opt out of the enterprise entirely. There is an unfortunate cultural norm within psychology (and, to be fair, many other fields) to demand that every research contribution end on a wholly positive or "constructive" note. This is an indefensible expectation that I

won't bother to indulge. In life, you often can't have what you want, no matter how hard you try. In such cases, I think it's better to recognize the situation for what it is sooner rather than later. The fact that a researcher is able to formulate a question in his or her head that seems sensible (e.g., "does ego depletion exist"?) doesn't mean that the question really *is* sensible. Moreover, even when the question is a sensible one to *ask* (in the sense that it's logically coherent and seems theoretically meaningful), it doesn't automatically follow that it's worth trying to obtain an empirical answer. In many research areas, if generalizability concerns were to be taken seriously, the level of effort required to obtain even minimally informative answers to seemingly interesting questions would likely so far exceed conventional standards that I suspect many academic psychologists would, if they were dispassionate about the matter, simply opt out. I see nothing wrong with such an outcome, and think it is a mistake to view a career in psychology (or any other academic field) as a higher calling of some sort.

Admittedly, the utility of this advice depends on one's career stage, skills, and interests. It should not be terribly surprising if few tenured professors are eager to admit (even to themselves) that they have, as Paul Meehl rather colorfully put it, "achieved some notoriety, tenure, economic security and the like by engaging, to speak bluntly, in a bunch of nothing" (Meehl, 1990b, p. 230). The situation is more favorable for graduate students and postdocs, who have much less to lose (and potentially much more to gain) by pursuing alternative careers. To be clear, I'm not suggesting that a career in academic psychology isn't a worthwhile pursuit *for anyone*; for many people, it remains an excellent choice. But I do think all psychologists, and early-career researchers in particular, owe it to themselves to spend some time carefully and dispassionately assessing the probability that the research they do is going to contribute meaningfully – even if only incrementally – to our collective ability either to understand the mind or to practically improve the human condition. There is no shame whatsoever in arriving at a negative answer, and the good news is that, for people who have managed to obtain a Ph.D. (or have the analytical skills to do so), career prospects outside of academia have arguably never been brighter.

6.2. Embrace qualitative analysis

A second approach one can take is to keep doing psychological research, but to largely abandon inferential statistical methods in favor of qualitative methods. This may seem like a radical prescription, but I contend that a good deal of what currently passes for empirical psychology is already best understood as insightful qualitative analysis trying to quietly pass for quantitative science. Careful consideration of the logical structure of a psychological theory often makes it clear that there is little point in subjecting the theory to quantitative analysis. Sometimes, this is because the theory appears logically incoherent, or is so vague as to make falsification via statistical procedures essentially impossible. Very often, however, it is because careful inspection reveals that the theory is actually *too* sensible. That is, its central postulates are so obviously true that there is nothing to be gained by subjecting it to further empirical tests – effectively constituting what Smedslund (1991) dubbed "pseudoempirical research."

To see what I mean, let's return to our running example of verbal overshadowing.

To judge by the accumulated literature (for reviews, see Meissner & Brigham, 2001; Meissner & Memon, 2002), the question of whether verbal overshadowing is or is not a "real" phenomenon seems to be taken quite seriously by many researchers. Yet it's straightforward to show that some phenomenon like verbal overshadowing must exist given even the most basic, uncontroversial facts about the human mind. Consider the following set of statements:

1. The human mind has a finite capacity to store information.
2. There is noise in the information-encoding process.
3. Different pieces of information will sometimes interfere with one another during decision-making – either because they directly conflict, or because they share common processing bottlenecks.

None of the above statements should be at all controversial, yet the conjunction of the three logically entails that there will be (many) situations in which something we could label verbal overshadowing will predictably occur. Suppose, we take the set of all situations in which a person witnesses, and encodes into memory, a crime taking place. In some subset of these cases, that person will later reconsider, and verbally re-encode, the events they observed. Because the encoding process is noisy, and conversion between different modalities is necessarily lossy, some details will be overemphasized, underemphasized, or otherwise distorted. And because different representations of the same event will conflict with one another, it is then guaranteed that there will be situations in which the verbal reconsideration of information at time 2 will lead a person to incorrectly ignore information they may have correctly encoded at time 1. We can call this *verbal overshadowing* if we like, but there is nothing about the core idea that requires any kind of empirical demonstration. So long as it's framed strictly in broad qualitative terms, the "theory" is trivially true; the only way it could be false is if at least one of the three statements listed above is false – which is almost impossible to imagine. (Note too, that the inverse of the theory is also trivially true: There must be many situations in which lossy re-encoding of information across modalities actually ends up being accidentally beneficial.)

To be clear, I am not suggesting that there's no point in quantitatively studying broad putative constructs like verbal overshadowing. On the contrary, if our goal is to develop models detailed enough to make useful real-world predictions, quantitative analysis may be indispensable. It would be difficult to make real-world predictions about when, where, and to what extent verbal overshadowing will manifest unless one has systematically studied and modeled the putative phenomenon under a broad range of conditions – including extensive variation of the perceptual stimuli, viewing conditions, rater incentives, timing parameters, and so on and so forth. But taking this quantitative objective seriously requires much larger and more complex datasets, experimental designs, and statistical models than have typically been deployed in most areas of psychology. As such, psychologists intent on working in "soft" domains who are unwilling to learn potentially challenging new modeling skills – or to spend months or years trying to meticulously address "minor" methodological concerns that presently barely rate any mention in papers – may need to accept that their research is, at root, qualitative in nature, and that the inferential statistics so often reported in soft psychology articles primarily serve as a ritual intended to convince one's

colleagues and/or one's self that something very scientific and important is taking place.

What would a qualitative psychology look like? In many subfields, almost nothing would change. The primary difference is that researchers would largely stop using inferential statistics, restricting themselves instead to descriptive statistics and qualitative discussion. Such a policy is not without precedent: in 2014, the journal *Basic and Applied Social Psychology* (BASP) banned the reporting of *p*-values from all submitted manuscripts (Trafimow, 2014; Trafimow & Marks, 2015). Although the move was greeted with derision by many scientists (Woolston, 2015), what is problematic about the BASP policy is, in my view, only that the abolition of inferential statistics was made mandatory. Framed as a strong recommendation that psychologists should avoid reporting inferential statistics that they often do not seem to understand, and that have no clear implications for our understanding of, or interaction with, the world, I think there would be much to like about the policy.

For many psychologists, fully embracing qualitative analysis would provide an explicit license to do what they are already most interested in doing – namely, exploring big ideas, generalizing conclusions far and wide, and moving swiftly from research question to research question. The primary cost would be the reputational one: In a world where most psychology papers are no longer accompanied by scientific-looking inferential statistics, journalists and policymakers would probably come knocking on our doors less often. I don't deny that this is a non-trivial cost, and I can understand why many researchers would be hesitant to pay such a toll. But such is life. I don't think it requires a terribly large amount of intellectual integrity to appreciate that one shouldn't portray one's self as a serious quantitative scientist unless one is actually willing to do the corresponding research.

Lest this attitude seem overly dismissive of qualitative approaches, it's worth noting that the core argument made in this paper is itself a qualitative one. I do not rely on inferential statistical results to support my conclusions, and all of the empirical data I quantitatively analyze are used strictly to illustrate general principles. Put differently, I am not making a claim of the form "87% of psychology articles draw conclusions that their data do not support"; I am observing that under modest assumptions that seem to me almost impossible to dispute in most areas of psychology (e.g., that the aggregate contribution of random variation in factors like experimental stimuli, task implementation, experimenter, site, and so on, is (1) large, and (2) almost never modeled), it is logically entailed that the conclusions researchers draw verbally will routinely deviate markedly from what the reported statistical analyses can strictly support. Researchers are, of course, free to object that this sweeping conclusion might not apply to *their* particular study, or that the argument would be more persuasive if accompanied by a numerical estimate of the magnitude of the problem in different areas.[10] But the mere fact that an argument is qualitative rather than quantitative in nature does not render it inferior or dismissible. On the contrary, as the verbal overshadowing example above illustrates, even a relatively elementary qualitative analysis can often provide more insightful answers to a question than a long series of ritualistic quantitative analyses. Therefore, I mean it sincerely when I say that an increased emphasis on qualitative considerations would be a welcome development in its own right in psychology, and should not be viewed as a consolation prize for studies that fail to report enough numbers.

6.3. Adopt better standards

The previous two suggestions are not a clumsy attempt at dark humor; I am firmly convinced that many academic psychologists would be better off either pursuing different careers, or explicitly acknowledging the fundamentally qualitative nature of their research (I lump myself into the former group much of the time, and this paper itself exemplifies the latter). For the remainder – that is, those who would like to approach their research from a more quantitatively defensible perspective – there are a number of practices that, if deployed widely, could greatly improve the quality and reliability of quantitative psychological inference.

6.3.1. Draw more conservative inferences

Perhaps the most obvious, and arguably easiest, solution to the generalizability problem is for authors to draw much more conservative inferences in their manuscripts – and in particular, to replace the sweeping generalizations pervasive in contemporary psychology with narrower conclusions that hew much more closely to the available data. Concretely, researchers should avoid extrapolating beyond the universe of observations implied by their experimental designs and statistical models without clearly indicating that they are engaging in speculation. Potentially relevant design factors that are impractical to measure or manipulate, but that conceptual considerations suggest are likely to have non-trivial effects (e.g., effects of stimuli, experimenter, research site, culture, etc.), should be identified and disclosed to the best of authors' ability. Papers should be given titles like "Transient manipulation of self-reported anger influences small hypothetical charitable donations," and not ones like "Hot head, warm heart: Anger increases economic charity." I strongly endorse the recent suggestion by Simons and colleagues that most manuscripts in psychology should include a *Constraints on Generality* statement that explicitly defines the boundaries of the universe of observations the authors believe their findings apply to (Simons, Shoda, & Lindsay, 2017) – as well as earlier statements to similar effects in other fields (e.g., sociology; Walker & Cohen, 1985).

Correspondingly, when researchers evaluate results reported by others, credit should only be given for what the empirical results of a study *actually* show – and not for what its authors claim they show. Continually emphasizing the importance of the distinction between verbal constructs and observable measurements would go a long way toward clarifying which existing findings are worth replicating and which are not. If researchers develop a habit of mentally reinterpreting a claim like "we provide evidence of ego depletion" as "we provide evidence that crossing out the letter $_e$ slightly decreases response accuracy on a subsequent Stroop task," I suspect that many findings would no longer seem important enough to warrant any kind of follow-up – at least, not until the original authors have conducted considerable additional research to demonstrate the generalizability of the claimed phenomenon.

6.3.2. Take descriptive research more seriously

Traditionally, purely descriptive research – where researchers seek to characterize and explore relationships between measured variables without imputing causal explanations or testing elaborate verbal theories – is looked down on in many areas of psychology. This stigma discourages modesty, inhibits careful characterization of phenomena, and often leads to premature and overconfident

efforts to assess simplistic theories that are hopelessly disconnected from the complexity of the real world (Cronbach, 1975; Rozin, 2001). I suspect it stems to a significant extent from a failure to recognize and internalize just how fragile many psychological phenomena truly are. Acknowledging the value of empirical studies that do nothing more than carefully describe the relationships between a bunch of variables under a wide range of conditions would go some ways toward atoning for our unreasonable obsession with oversimplified causal explanations.

We know that a large-scale shift in expectations regarding the utility of careful descriptive research is possible, because other fields have undergone such a transition to varying extents. Perhaps most notably, in statistical genetics, the small-sample candidate gene studies that made regular headlines in the 1990s (e.g., Ebstein et al., 1996; Lesch et al., 1996) – virtually all of which later turned out to be spurious (Chabris et al., 2012; Colhoun, McKeigue, & Davey Smith, 2003; Sullivan, 2007), and were motivated by elegant theoretical hypotheses that seem laughably simplistic in hindsight – have all but disappeared in favor of massive genome-wide association studies (GWASs) involving hundreds of thousands of subjects (Nagel et al., 2018; Savage et al., 2018; Wray et al., 2018). The latter are now considered the gold standard even in cases where they do little more than descriptively identify novel statistical associations between gene variants and behavior. In much of statistical genetics, at least, researchers seem to have accepted that the world is causally complicated, and attempting to obtain a reasonable descriptive characterization of some small part of it is a perfectly valid reason to conduct large, expensive empirical studies.

6.3.3. Fit more expansive statistical models

To the degree that authors intend for their conclusions to generalize over populations of stimuli, tasks, experimenters, and other such factors, they should develop a habit of fitting more expansive statistical models. As noted earlier, nearly all statistical analyses of multisubject data in psychology treat subject as a varying effect. The same treatment should be accorded to other design factors that researchers intend to generalize over and that vary controllably or naturally in one's study. Of course, inclusion of additional random effects is only one of many potential avenues for sensible model expansion (Draper, 1995; Gelman & Shalizi, 2013).[11] The good news is that improvements in statistical computing over the past few years have made it substantially easier for researchers to fit arbitrarily complex mixed-effects models within both Bayesian and frequentist frameworks. Models that were once intractable for most researchers to fit because of either mathematical or computational limitations can now often be easily specified and executed on modern laptops using mixed-effects packages (e.g., lmer or MixedModels.jl; Bates, Maechler, Bolker, Walker, Christensen, Singmann, et al., 2014) or probabilistic programing frameworks (e.g., Stan or PyMC; Carpenter et al., 2017; Salvatier, Wiecki, & Fonnesbeck, 2016).

This recommendation conveniently sidesteps the question of *which* varying factors researchers should choose to focus on. A number of commentators on earlier drafts of this paper have suggested that the general prescription to fit bigger models, whereas technically reasonable, is too vague to be helpful. I am sympathetic to this concern, but nevertheless think that attempting to make generic statements about the relative importance of different sources of variation in "typical" psychology studies would be a mistake. There are two reasons for this. First, I see little reason to think that any brief domain-general summary of the relative magnitudes of different variance components would have much utility for almost any individual study. How important is it to consider the role of different task operationalizations? Do cross-cultural differences have a small or large impact on observed effect sizes? And what about experimenter effects, how big are those? The only answer one can give to such questions that is both honest and concise is "it depends."

Second, the sense of discomfort some readers might feel at the realization that they don't know what to do next is, in my view, a feature, not a bug. It *should* bother researchers to discover that they don't have a good sense of what the major sources of variance are in the data they routinely work with. What does it say about a researcher's ability to update their belief in a hypothesis if they cannot even roughly state the conditions under which the obtained statistical results would or would not constitute an adequate test of the hypothesis? I would not want to give researchers the impression that there is some generic list of factors one can rely on here; there is simply no substitute for careful and critical consideration of the data-generating processes likely to underlie each individual effect of interest.

6.3.4. Design with variation in mind

In most areas of psychology, there is a long-dominant tradition of trying to construct randomized experiments that are as tightly controlled as possible – even at the cost of decreased generalizability. Although calls for researchers to emphasize the opposite side of the precision-generalization trade-off – that is, to embrace naturalistic, ecologically valid designs that embrace variability – have a long history in psychology (Brunswik, 1947; Cronbach, 1975), they have intensified considerably in recent years. For example, in neuroimaging, researchers are increasingly fitting sophisticated models to naturalistic stimuli such as coherent narratives or movies (Hamilton & Huth, 2018; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Huth, Nishimoto, Vu, & Gallant, 2012; Spiers & Maguire, 2007). In psycholinguistics, large-scale analyses involving databases of thousands of words and subjects have superseded traditional small-$_n$ factorial studies for many applications (Balota, Yap, Hutchison, & Cortese, 2012; Keuleers & Balota, 2015). Even in domains where many effects traditionally display little sensitivity to context, some researchers have advocated for analysis strategies that emphasize variability. For example, Baribault and colleagues (2018) randomly varied 16 different experimental factors in a large multisite replication (6 sites, 346 subjects, and nearly 5,000 "microexperiments") of a subliminal priming study (Reuss, Kiesel, & Kunde, 2015). The "radical randomization" strategy the authors adopted allowed them to draw much stronger conclusions about the generalizability of the priming effect (or lack thereof) than would have otherwise been possible.

The deliberate introduction of variance into one's studies can also be construed as a more principled version of the *conceptual replication* strategy already common in many areas of psychology. In both cases, researchers seek to determine the extent to which an effect generalizes across the levels of one or more secondary design factors. The key difference is that traditional conceptual replications do not lend themselves well to a coherent modeling strategy: When authors present a series of discrete conceptual replications in studies two through $_N$ of a manuscript, it is rarely obvious how one can combine the results to obtain a meaningful estimate of the robustness or generalizability of the common effect. By contrast, explicitly modeling the varying factors as components of a single overarching design

makes it clear what the putative relationship between different measurements is, and enables stronger quantitative inferences to be drawn.

Naturally, variation-enhancing designs come at a cost: they will often demand greater resources than conventional approaches that seek to minimize extraneous variation. But, if authors intend for their conclusions to hold independently of variation in uninteresting factors, and to generalize to broad classes of situations, there is no good substitute for studies whose designs make a serious effort to respect and capture the complexity of real-world phenomena. Large-scale, collaborative projects of the kind pioneered in RRRs (Simons et al., 2014) and recent initiatives such as the Psychology Accelerator (Moshontz et al., 2018) are arguably the natural venue for such an approach – but, to maximize their utility, the substantial resources they command must be used to directly measure and model variability rather than minimizing and ignoring it.

6.3.5. Emphasize variance estimates

An important and underappreciated secondary consequence of the widespread disregard for generalizability is that researchers in many areas of psychology rarely have good data – or even just strong intuitions – about the relative importances of different sources of variance. One way to mitigate this problem is to promote analytical approaches that emphasize the estimation of variance components rather than focusing solely on point estimates. For primary research studies, Generalizability Theory (Brennan, 1992; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991) provides a well-developed (and underused) framework for computing and applying such estimates. At the secondary level, meta-analysts could similarly work to quantify the magnitudes of different variance components – either by meta-analyzing reported within-study variance estimates, or by meta-analytically computing between-study variance components for different factors. Such approaches could provide researchers with critically important background estimates of the extent to which a new finding reported in a particular literature should be expected to generalize to different samples of subjects, stimuli, experimenters, research sites, and so on. Notably, such estimates would be valuable irrespective of the presence or absence of a main effect of putative interest. For example, even if the accumulated empirical literature is too feeble to allow us to estimate anything approximating a single overall *universe score* for ego depletion, it would still be extremely helpful when planning a new study to know roughly how much of the observed variation in the existing pool of studies is because of differences in stimuli, subjects, tasks, and so on.

6.3.6. Make riskier predictions

There is an important sense in which most of the other recommendations made in this section could be obviated simply by making theoretical predictions that assume a high degree of theoretical risk. I have approached the problem of generalizability largely from a statistical perspective, but there is a deep connection between the present concerns and a long tradition of philosophical commentary focusing on the logical relationship (or lack thereof) between theoretical hypotheses and operational or statistical ones.

Perhaps, the best exposition of such ideas is found in the seminal study of Paul Meehl, who, beginning in the 1960s, argued compellingly that many of the methodological and statistical practices routinely applied by psychologists and other social scientists are logically fallacious (e.g., Meehl, 1967, 1978, 1990b). Meehl's thinking was extremely nuanced, but a recurring theme in his study is the observation that most hypothesis tests in psychology commit the logical fallacy of affirming the consequent. A theory $_T$ makes a prediction $_P$, and when researchers obtain data consistent with $_P$, they then happily conclude that $_T$ is corroborated. In reality, the confirmation of $_P$ provides no meaningful support for $_T$ unless the prediction was relatively specific to $_T$ – that is, there are no readily available alternative theories $T_1^0 \ldots T_k^0$ that also predict $_P$. Unfortunately, in most domains of psychology, there are pervasive and typically very plausible competing explanations for almost every finding (Cohen, 2016; Lykken, 1968; Meehl, 1967, 1986).

The solution to this problem is, in principle, simple: Researchers should strive to develop theories that generate risky predictions (Meehl, 1997; Meehl, 1967, 1990a; Popper, 2014) – or, in the terminology popularized by Deborah Mayo, should subject their theories to *severe tests* (Mayo, 1991, 2018). The canonical way to accomplish this is to derive from one's theory some series of predictions – typically, but not necessarily, quantitative in nature – sufficiently specific to that theory that they are inconsistent with, or at least extremely implausible under, other accounts. As Meehl put it:

> If my meteorological theory successfully predicts that it will rain sometime next April, and that prediction pans out, the scientific community will not be much impressed. If my theory enables me to correctly predict which of 5 days in April it rains, they will be more impressed. And if I predict how many millimeters of rainfall there will be on each of these 5 days, they will begin to take my theory very seriously indeed (Meehl, 1990a, p. 110).

The ability to generate and corroborate a truly risky prediction strongly implies that a researcher must already have a decent working model (even if only implicitly) of most of the contextual factors that could potentially affect a dependent variable. If a social psychologist was capable of directly deriving from a theory of verbal overshadowing the prediction that target recognition should decrease $1.7 \pm 0.04\%$ in condition A relative to condition B in a given experiment, concerns about the generalizability of the theory would dramatically lessen, as there would rarely be a plausible alternative explanation for such precision other than that the theories in question were indeed accurately capturing something important about the way the world works.

In practice, it's clearly wishful thinking to demand this sort of precision in most areas of psychology (potential exceptions include, e.g., parts of psychophysics, mathematical cognitive psychology, and behavioral genetics). The very fact that most of the phenomena psychologists study are enormously complex, and admit a vast array of causal influences in even the most artificially constrained laboratory situations, likely precludes the production of quantitative models with anything close to the level of precision one routinely observes in the natural sciences. This does not mean, however, that vague directional predictions are the best we can expect from psychologists. There are a number of strategies that researchers in such fields could adopt that would still represent at least a modest improvement over the status quo (for discussion, see Gigerenzer, 2017; Lilienfeld, 2004; Meehl, 1990a; Roberts & Pashler, 2000). For example, researchers could use equivalence tests (Lakens, 2017); predict specific orderings of discrete observations; test against compound nulls that require the conjunctive rejection of many independent directional

predictions; and develop formal mathematical models that posit non-trivial functional forms between the input and output variables (Marewski & Olsson, 2009; Smaldino, 2017). Although it is probably unrealistic to expect truly severe tests to become the norm in most fields of psychology, severity is an ideal worth keeping perpetually in mind when designing studies – if only as a natural guard against undue optimism.

6.3.7. Focus on practical predictive utility

An alternative and arguably more pragmatic way to think about the role of prediction in psychology is to focus not on the theoretical risk implied by a prediction, but on its practical utility. Here, the core idea is to view psychological theories or models not so much as statements about how the human mind *actually* operates, but as convenient approximations that can help us intervene on the world in useful ways (Breiman, 2001; Hofman, Sharma, & Watts, 2017; Shmueli, 2010; Yarkoni & Westfall, 2017). For example, instead of asking the question *does verbal overshadowing exist*?, we might instead ask: *Can we train a statistical model that allows us to meaningfully predict people's behaviors in a set of situations that superficially seem to involve verbal overshadowing*? The latter framing places emphasis primarily on what a model is able to *do* for us rather than on its implied theoretical or ontological commitments.

One major advantage of an applied predictive focus is that it naturally draws attention to objective metrics of performance that can be easier to measure and evaluate than the relatively abstract, and often vague, theoretical postulates of psychological theories. A strong emphasis on objective, communal measures of model performance has been a key driver of rapid recent progress in the field of machine learning (Jordan & Mitchell, 2015; LeCun, Bengio, & Hinton, 2015; Russakovsky et al., 2015) – including numerous successes in domains such as object recognition and natural language translation that arguably already fall within the purview of psychology and cognitive science. A focus on applied prediction would also naturally encourage greater use of large samples, as well as of cross-validation techniques that can minimize overfitting and provide alternative ways of assessing generalizability outside of the traditional inferential statistical framework. Admittedly, a large-scale shift toward instrumentalism of this kind would break with a century-long tradition of explanation and theoretical understanding within psychology; however, as I have argued elsewhere (Yarkoni & Westfall, 2017), there are good reasons to believe that psychology would emerge as a healthier, more reliable discipline as a result.

7. Conclusion

Most contemporary psychologists view the use of inferential statistical tests as an integral part of the discipline's methodology. The ubiquitous reliance on statistical inference is the source of much of the perceived objectivity and rigor of modern psychology – the very thing that, in many people's eyes, makes it a quantitative science. I have argued that, for most research questions in most areas of psychology, this perception is illusory. Closer examination reveals that the inferential statistics reported in psychology articles typically have only a tenuous correspondence to the verbal claims they are intended to support. The overarching conclusion is that many fields of psychology currently operate under a kind of collective self-deception, using a thin sheen of quantitative rigor to mask inferences that remain, at their core, almost entirely qualitative.

Such concerns are not new, of course. Commentators have long pointed out that, viewed dispassionately, an enormous amount of statistical inference in psychology (and, to be fair, other sciences) has a decidedly ritualistic character: rather than improving the quality of scientific inference, the use of universalized testing procedures serves mainly to increase practitioners' subjective confidence in broad verbal assertions that would otherwise be difficult to defend on logical grounds (e.g., Gelman, 2016; Gigerenzer, 2004; Gigerenzer & Marewski, 2015; Meehl, 1967, 1990b; Tong, 2019). What I have tried to emphasize in the present treatment is that such critiques are not, as many psychologists would like to believe, pedantic worries about edge cases that one can safely ignore most of the time. The problems in question are fundamental, and follow directly from foundational assumptions of our most widely used statistical models. The central point is that the degree of support a statistical analysis lends to a verbal proposition derives not just from some critical number that the analysis does or doesn't pop out (e.g., $_p < 0.05$), but also (and really, *primarily*), from the ability of the statistical model to implicitly define a universe that matches the one defined by the verbal proposition.

When the two diverge markedly – as I have argued is extremely common in psychology – one is left with a difficult choice to make. One possibility is to accept the force of the challenge and adjust one's standard operating procedures accordingly – by moderating one's verbal claims, narrowing the scope of one's research program, focusing on making practically useful predictions, and so on. This path is effort-intensive and incurs a high risk that the results one produces post-remediation will, at least superficially, seem less impressive than the ones that came before. But it is the intellectually honest road, and has the secondary benefit of reducing the probability of making unreasonably broad claims that are unlikely to stand the test of time.

The alternative is to simply brush off these concerns, recommit one's self to the same set of procedures that have led to prior success by at least *some* measures (papers published, awards received, etc.), and then carry on with business as usual. No additional effort is required here; no new intellectual or occupational risk is assumed. The main cost is that one must live with the knowledge that many of the statistical quantities one routinely reports in one's papers are essentially just an elaborate rhetorical ruse used to mathematize people into believing claims they would otherwise find logically unsound.

I don't pretend to think this is an easy choice. I have little doubt that the vast majority of researchers have good intentions, and genuinely want to do research that helps increase understanding of human behavior and improve the quality of people's lives. I am also sympathetic to objections that it's not fair to expect individual researchers to proactively hold themselves to a higher standard than the surrounding community, knowing full well that a likely cost of doing the right thing is that one's research may become more difficult to pursue, less exciting, and less well received by others. Unfortunately, the world we live in isn't always fair. I don't think anyone should be judged very harshly for finding major course correction too difficult an undertaking after spending years immersed in an intellectual tradition that encourages rampant overgeneralization. But the decision to stay the course should at least be an informed one: Researchers who opt to ignore the bad news should recognize that, in the long term,

such an attitude hurts the credibility both of their own research program and of the broader profession they have chosen. One is always free to pretend that small *p*-values obtained from extremely narrow statistical operationalizations can provide an adequate basis for sweeping verbal inferences about complex psychological constructs. But no one else – not one's peers, not one's funders, not the public, and certainly not the long-term scientific record – is obligated to honor the charade.

Acknowledgments. This paper was a labor of pain that took an inexplicably long time to produce. It owes an important debt to Jake Westfall for valuable discussions and comments – including the idea of fitting the "unmeasured variances" model illustrated in Figure 2 – and also benefited from conversations with dozens of other people (many of them on Twitter, because it turns out you can say a surprisingly large amount in hundreds of 280-character tweets).

Financial support. This study was supported by NIH awards R01MH096906 and R01MH109682.

Conflict of interest. None.

Notes

1. I avoid the conventional habit of describing psychological constructs as latent variables, as such language is often taken to imply a realist philosophical stance toward theoretical entities (e.g., Borsboom, Mellenbergh, & van Heerden, 2003). For present purposes, it's irrelevant whether one thinks psychological constructs objectively exist in some latent or platonic realm, or are merely pragmatic fictions.
2. The precise effect of failing to include random factors depends on a number of considerations, including the amount of variance between vs. within the random effects, the covariance with other variables, and the effective sample sizes of different factors. But in most real-world settings, the inclusion of random effects will lead to (often much) larger uncertainty estimates and smaller inferential test statistics.
3. To keep things simple, I ignore the question of how one ought to decide whether or not to include both random slopes and random intercepts (for discussion, see Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The goal, here, is simply to elucidate the distinction between fixed and random effects.
4. Note that in econometrics, the term *fixed effect* has a narrower meaning, and refers specifically to a group mean parameter (rather than just any predictor variable) modeled as non-random.
5. A reasonable argument could be made that since no experimental context is ever *exactly* the same across two measurement occasions, in a technical sense, no design factor is ever truly fixed. Readers who are sympathetic to such an argument (as I also am) should remember Box's dictum that "all models are false, but some are useful," and are invited to construe the choice between fixed and random effects as a purely pragmatic one that amounts to deciding which of two idealizations better approximates reality.
6. That even small differences in such factors can have large impacts on the outcome is clear from the Alogna et al. (2014) study itself: because of an error in the timing of different components of the procedure, Alogna et al. actually conducted *two* large replication studies. They observed a markedly stronger effect when the experimental task was delayed by 20 min than when it immediately followed the video.
7. For example, perhaps participants in Alogna et al.'s experimental condition felt greater pressure to produce the correct answer (having previously spent several minutes describing their perceptions), and it was the stress rather than the treatment *per se* that resulted in poorer performance. Or, perhaps the effect had nothing at all to do with the treatment condition, and instead reflected a poor choice of control condition (say, because naming countries and capitals incidentally activates helpful memory consolidation processes). And so on and so forth. (A skeptic might object that each such explanation is individually not as plausible as the verbal overshadowing account, but this misses the point: safely generalizing the results of the narrow Schooler and Engstler-Schooler (1990) design to the broad construct of verbal overshadowing implies that one can rule out the influence of *all* other confounds in the aggregate – and reality is not under any obligation to only manifest sparse causal relationships that researchers find intuitive!)
8. The model differs in that I fit a single mixed-effects linear probability model with random intercepts and slopes for sites, whereas Alogna et al. first computed the mean difference in response accuracy between conditions for each site, and then performed a random-effects meta-analysis (note that a logistic regression model would be appropriate here given the binary outcome, but I opted for the linear model for the sake of consistency with Alogna et al. and simplicity of presentation). The data differ because (a) some sites' datasets were not publicly available, (b) I made no attempt to adhere closely to the reported preprocessing procedures (e.g., inclusion/exclusion criteria), and (c) I used only the data from the (more successful) second RRR reported in the paper. All data and code used in the analyses reported here are available at https://github.com/tyarkoni/generalizability.
9. We should probably be cautious in drawing even this narrow conclusion, however, because the experimental procedure in question could very well be producing the observed effect because of idiosyncratic and uninteresting properties, and not because it induces verbal overshadowing *per se*.
10. For what it's worth, it's unclear how much utility global quantitative estimates of this kind could actually have given the enormous variation across studies, and the relative ease of obtaining directly relevant local estimates. Individual researchers who want to know whether or not it is safe to assume zero stimulus, experimenter, or task effects in their statistical models do not have to wait for someone else to conduct a comprehensive variance-partitioning meta-analysis in their general domain; they can simply calculate the variance over such factors in their own prior datasets!
11. In a sense, the very idea of a random effect is just a convenient fiction – effectively, a placeholder for a large number of hypothetical fixed variables (or functions thereof) that we presently do not know how to write, or lack the capacity to measure and/or estimate.

References

Acosta, A., Adams, Jr., R. B., Albohn, D. N., Allard, E. S., Beek, T., Benning, S. D., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š, Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59* (4), 390–412.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (pp. 90–115). Psychology Press.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2607–2612.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Krivitsky, P. N. (2014). Lme4: Linear mixed-effects models using eigen and S4. *R Package Version*, *1*(7), 1–23.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., & Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Bergelson, E., Bergmann, C., Byers-Heinlein, K., Cristia, A., Cusack, R., & Dyck, K., … (2017). Quantifying sources of variability in infancy research using the infant-directed speech preference.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Brunswik, E. (1947). Systematic and representative design of psychological experiments. In *Proceedings of the Berkeley symposium on mathematical statistics and probability* (pp. 143–202).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., … Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, *23*(11), 1314–1323.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š, … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Cohen, J. (2016). The earth is round ($p < 0.05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 69–82). Routledge.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*(1), 219–226.

Colhoun, H. M., McKeigue, P. M., & Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet (London, England)*, *361*(9360), 865–872.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*(4), 907–949.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science*, *284*(5420), 1670–1672.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology*, *59*(1), 20–26.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Mathematical and Statistical Psychology*, *16*(2), 137–163.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45–70.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., … Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. *Nature Genetics*, *12*(1), 78–80.

Eerland, A. S., Magliano, A. M., Zwaan, J. P., Arnal, R. A., Aucoin, J. D., & Crocker, P. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171.

Feynman, R. P. (1974). Cargo cult science. *Engineering Sciences*, *37*(7), 10–13.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*(2), 632–643.

Gelman, A. (2016). The problems with *p*-values are not just with *p*-values. *The American Statistician*, *70*(supplemental material to the ASA statement on p-values and statistical significance), 10.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis. *Downloaded January*, 1–17.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G. (2017). A theory integration program. *Decision*, *4*(3), 133.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, *11*(3), 385–398.

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, *19*(5), 640–648.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178–206.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), e105825.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., … Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science*, *274*(5292), 1527–1531.

Lilienfeld, S. O. (2004). Taking theoretical risks in a world of directional predictions. *Applied and Preventive Psychology*, *11*(1), 47–51.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 49–60.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*(4), 523–552.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(Suppl. 1), 235–245.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Erlbaum.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806.

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–338). University of Chicago Press.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*(6), 603–616.

Meissner, C. A., & Memon, A. (2002). Verbal overshadowing: A special issue exploring theoretical and applied issues. *Applied Cognitive Psychology*, *16*(8), 869–872.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). Psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., Bryois, J., … Posthuma, D. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics*, *50*(7), 920–927.

O'Leary-Kelly, S. W., & Vokurka, R. J. (1998). The empirical assessment of construct validity. *Journal of Operations Management*, *16*(4), 387–405.
Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.
Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.
Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *T* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.
Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.
Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55.
Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., … Posthuma, D. (2018). Genome-wide association metaanalysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, *50*(7), 912–919.
Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE.
Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.
Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.
Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555.
Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.
Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). Routledge.
Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.
Smedslund, J. (1991). The pseudoempirical in psychology and the case for psychologic. *Psychological Inquiry*, *2*(4), 325–338.
Spiers, H. J., & Maguire, E. A. (2007). Decoding human brain activity during real-world experiences. *Trends in Cognitive Sciences*, *11*(8), 356–365.
Steckler, A., McLeroy, K. R., Goodman, R. M., Bird, S. T., & McCormick, L. (1992). Toward integrating qualitative and quantitative methods: An introduction. *Health Education Quarterly*, *19*(1), 1–8.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.
Sullivan, P. F. (2007). Spurious genetic associations. *Biological Psychiatry*, *61*(10), 1121–1126.
Tong, C. (2019). Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, *73*(Suppl. 1), 246–261.
Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, *36*(1), 1–2.
Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2.
Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6454–6459.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.
Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., & Dorow, J., … (2003). Different data from different labs: Lessons from studies of gene–environment interaction. *Journal of Neurobiology*, *54*(1), 283–311.
Walker, H. A., & Cohen, B. P. (1985). Scope statements: Imperatives for evaluating theory. *American Sociological Review*, *50*, 288–301.
Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, *1*, 23.
Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209.
Woolston, C. (2015). Psychology journal bans *P* values. *Nature News*, *519*(7541), 9.
Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., & Abdellaoui, A., … Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*(5), 668–681.
Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power-commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Open Peer Commentary

There is no psychology without inferential statistics

Shilaan Alzahawi and Benoît Monin

Graduate School of Business, Stanford University, Knight Management Center, Stanford, CA 94305, USA
shilaan@stanford.edu; https://shilaan.rbind.io
monin@stanford.edu; https://monin.people.stanford.edu

doi:10.1017/S0140525X2100056X, e2

Abstract

Quantification has been constitutive of psychology since its inception and is core to its scientific status. The adoption of qualitative methods eschewing inferential statistics is therefore unlikely to obtain. Rather than discarding useful tools because of improper use, we recommend highlighting how inferential statistics can be more thoughtfully applied.

> Why take the trouble to compute *p*-values, Bayes Factors, or confidence intervals when evaluating qualitative theoretical claims? Why don't psychologists simply look at the world around them, think deeply for a while, and then state – again in qualitative terms – what they think they have learned?
>
> (Yarkoni, sect. 1, para. 1).

Yarkoni convincingly argues that psychology is filled with studies that spuriously claim support for fixed effects while these findings often result from – and will not generalize beyond – the specific stimuli, measures, or manipulations used. One of Yarkoni's proposed solutions is that one could "largely abandon inferential statistical methods in favor of qualitative methods" (sect. 6.2, para. 1), and he argues "sincerely" that "an increased emphasis on qualitative considerations would be a welcome development in its own right in psychology" (sect. 6.2, para. 7).[1]

In this commentary, we argue that Yarkoni's call for psychology to embrace qualitative methods is likely to fall on deaf ears, because psychology has treated quantitative criteria as touchstones of psychological progress since its very inception. From the brass instruments of early psychological laboratories to the fetishization of the *p*-value in recent decades, quantitative methods have been central to psychology's self-image and self-presentation as an objective, scientific discipline.

From its onset, quantitative methods have been constitutive of psychology because they allowed the nascent discipline to contrast itself from other inquiries into the human mind. In the early years of the discipline, psychology's object of study had strong overlap with existing academic fields such as philosophy, but also with non-academic sources of knowledge like commonsensical beliefs, quacks, and pseudo-scientists. Because of this overlap in its subject matter, it was crucial for psychology to distinguish itself through its methods. The adoption of statistics – with their veneer of rigor and precision – helped establish psychology as a distinct academic discipline (Coon, 1993, pp. 762–763), and in turn afforded psychologists scientific status, separate resources, and university positions (Ash, 1992, p. 198). With time, rules related to quantification even came to define psychology's subject matter: phenomena resisting quantification were excluded from psychology's inquiry, and remaining phenomena were redefined to be quantifiable (Hornstein, 1988, pp. 21–22).

Another reason for the eager adoption of quantitative methods was that agreement over methods helped paper over wide disagreement over theory. Early psychology struggled to reach consensus on theory and subject matter. Rigid rules of quantification offered an attractive strategy to ignore these theoretical difficulties and provided a common language to unify and regulate an otherwise disparate field (Danziger, 1990, pp. 148–153).

As methods evolved, they promised ever greater objectivity. By mid-century, the institutionalization of inferential statistics further unified psychology at the methodological level (Gigerenzer & Murray, 1987, pp. 19–20). Inferential statistics, and particularly *p*-values, were welcomed as seemingly theory-neutral devices that mechanized the acquisition of knowledge while eliminating the need for personal judgment.

In short, quantitative methods are core to psychology's social and scientific status. They have helped – and continue to help – psychology establish itself as an independent discipline, negotiate its boundaries with neighboring disciplines, demarcate itself from pseudo-science and lay observation, and compete for scarce resources. *p*-values and other inferential statistical methods are more than tools for the pursuit of theory. Their value as signals of academic status and scientific rigor is what makes them so appealing to psychology as a discipline and to individual psychologists eager for status in academia and beyond.

Given how much psychology relies on inferential statistics for both its self-definition and its social status, we conclude that we are unlikely to see any significant uptake of qualitative methods in psychology despite Yarkoni's exhortation. Without inferential statistics, how does psychology justify its existence, its distinctiveness from other disciplines studying the human mind, and its claims to scientific status? Suggesting that psychologists abandon inferential statistics is like asking them "to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them" (Wasserstein, Schirm, & Lazar, 2019, p. 1).

The way forward is not to discard inferential statistics altogether; it is to improve the ways psychologists draw statistical inferences. Yarkoni provides a useful step in that direction. Rather than discarding the tool altogether on the basis of its improper use by some, we recommend highlighting how inferential statistics can be more thoughtfully applied. While debates rage among statisticians and psychologists about correct statistical inference, most agree that we should reject mindless, mechanical inference – using statistics as "a rote, mechanical procedure for turning data into conclusions" (sect. 1, para. 5) – and that we should instead embrace the importance of human judgment in statistical thinking.

Inferential statistics do not obviate the need for human judgment. Instead, active justification is essential at every stage of the process – for instance, when constructing hypotheses, selecting statistical models, and, to Yarkoni's main thesis, when making claims about generalizability. As a discipline, we stand to benefit tremendously from the creation and popularization of practical tools and resources helping psychologists actively make and justify these decisions. Therefore, we call not for the adoption of qualitative methods, but for a new era of education and thoughtful application of inferential statistics to draw more accurate – if more modest – conclusions from our data.

We can use inferential statistics better, but not abandon them. There is no psychology without inferential statistics.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

Note

1. We follow his idiosyncratic nomenclature below and define "qualitative" as any methods that do not involve inferential statistics.

References

Ash, M. G. (1992). Historicizing mind science: Discourse, practice, subjectivity. *Science in Context*, *5*(2), 193–207. https://doi.org/10.1017/S0269889700001150.

Coon, D. J. (1993). Standardizing the subject: Experimental psychologists, introspection, and the quest for a technoscientific ideal. *Technology and Culture*, *34*(4), 757–783. https://doi.org/10.2307/3106414.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511524059.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Lawrence Erlbaum Associates, Inc.

Hornstein, G. A. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J. G. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 1–34). Yale University Press.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician* *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Random effects won't solve the problem of generalizability

Adam Bear[a] and Jonathan Phillips[b]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138, USA and [b]Program in Cognitive Science, Dartmouth College, Hanover, NH 03755, USA.
adambear@fas.harvard.edu; https://adambear.me
jonathan.s.phillips@dartmouth.edu;
https://www.dartmouth.edu/~phillab/phillips.html

doi:10.1017/S0140525X2100011X, e3

Abstract

Yarkoni argues that researchers making broad inferences often use impoverished statistical models that fail to include important sources of variation as random effects. We argue, however, that for many common study designs, random effects are inappropriate and insufficient to draw general inferences, as the source of variation is not random, but systematic.

Yarkoni compellingly argues that researchers often neglect important sources of variation in their statistical models. One of the most important sources of variation that often goes unmodeled is the experimental stimuli that researchers select (sect. 3.1). Yarkoni encourages researchers to statistically model stimuli as a random factor in a mixed-effects model. While this suggestion will no doubt improve generalizability for certain types of psychological studies, it is inadequate in many other cases.

Modeling stimuli as a random factor introduces a key assumption about the process by which the stimuli were generated – an assumption that, in many experiments, is almost certainly false. For stimuli to count as "random," the source of variation must actually be random. That is, the stimuli are assumed to be random draws from a (usually) normal distribution that mimics the true distribution of stimuli to which the researchers want to generalize. Because this sampling distribution is assumed to be centered on the true average effect size, only the variance around this effect size is estimated. As Yarkoni shows, when the model estimates more variance, the true average effect is less certain, and it is more difficult to generalize beyond the particular set of stimuli used in the experiment.

In a wide range of cases, this assumption of random sampling is suspect. Consider a study that investigates whether disgusting immoral actions elicit increased moral reprobation. Suppose researchers generate a set of scenarios involving immoral actions, some of which are disgusting and others of which are not; collect moral judgments from a large sample of online participants; and – having read Yarkoni's article – model both subjects and stimuli as random factors in their analysis. If the mixed-effects model yields a highly significant p-value for the disgustingness of the action, is the general conclusion that disgusting moral violations are judged (by WEIRD [western, educated, industrialized, rich, and democratic] people) to be morally worse than non-disgusting ones warranted?

Probably not. The researchers created their stimuli with a hypothesis in mind and were introspectively aware of which stimuli would elicit stronger or weaker moral judgments. As a result, even if the researchers intend to create a fair test, they will almost certainly be disinclined – consciously or unconsciously – to select stimuli that are unlikely to provide support for their hypothesis. In other words, the stimuli that the researchers chose to include in the study were not random draws from a representative population of moral violations, but were biased to favor a particular conclusion.

Indeed, a study by Strickland and Suben (2012) provides a real-world demonstration of how this can happen, albeit in a somewhat exaggerated setting. Groups of undergraduates were assigned the task of creating stimuli to test specific hypotheses from experimental philosophy, but different groups were given contradictory hypotheses. The different groups generated systematically different stimuli, which in turn influenced whether, and to what extent, they observed a statistically significant effect.

There is a further problem with modeling stimuli as a random factor when the stimuli are generated nonrandomly. If researchers are systematically selecting stimuli that tend to favor their hypothesis, the model will tend to underestimate the true amount of variation in the effect size across stimuli. Concretely, imagine that, in the true population of stimuli that the experimenter wants to generalize to, the effect size is normally distributed around 0. Yet the researchers' directional hypothesis motivates them to systematically sample stimuli from the right tail of this distribution. The variance in the effect size of this truncated distribution will be substantially smaller than the variance of the true distribution – barely more than a third of the size. Indeed, even if the experimenters have only a weak bias to *avoid* sampling stimuli whose effect sizes are more than a standard deviation in the *opposite* direction of their hypothesis, the variance of the resulting distribution will be less than two-thirds of the true population variance. Thus, when the stimuli are sampled with bias, a random-effects model will almost certainly underestimate how much the effect size varies across stimuli and, in turn, provide overly narrow confidence intervals around an already biased estimate.

The problems that we lay out here are, in principle, quite difficult to solve, as they cannot be corrected by a simple tweak to a statistical model. Even more troubling is the fact that in many cases, there is no obvious way of determining what even is the "true" population of stimuli that one should generalize to. For example, is the effect of disgust on moral judgment meant to generalize to all *possible* actions in all possible scenarios? All *actual* morally relevant actions? Only some particular subset of salient moral violations? There seems to be no easy resolution to this question and, in turn, no easy way to know whether the stimuli represent a "biased" sample from the underlying "true" distribution.

Although this problem may seem intractable – and has even led us to question some of our own work – certain steps can be taken to mitigate it. For example, as Yarkoni suggests, experimenters can try to sample stimuli directly from real-world corpora (e.g., a court database of crimes). However, this is often laborious and impractical and, in fact, may suffer from its own biases (e.g., crimes may not be the category of immoral actions to which the researchers want to generalize). Alternatively, as Strickland and Suben (2012) suggest, researchers may recruit naive assistants or Mechanical Turk workers to generate stimuli without knowledge of the hypothesis. Finally, researchers could fight fire with fire by starting adversarial collaborations in which teams of researchers with the opposite hypotheses generate their own stimuli. If adversarial teams find effects of approximately equal magnitude in opposite directions, the original effect was likely due to experimenter bias. If not, the researchers should be more confident that their effect is generalizable to a broader stimulus set, even if these researchers cannot precisely define what that set is.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, *3*, 457–467.

Causal analysis as a bridge between qualitative and quantitative research

Rosemary Blersch[a,b], Neil Franchuk[a], Miranda Lucas[a], Christina M. Nord[a,b], Stephanie Varsanyi[a,b] and Tyler R. Bonnell[a,b]

[a]Barrett-Henzi Lab, Department of Psychology, University of Lethbridge, Lethbridge, AB, Canada and [b]Applied Behavioural Ecology and Ecosystems Research Unit, University of South Africa, Pretoria, South Africa, Lethbridge, AB T1K 3M4, Canada
blerschr@uleth.ca, neil.franchuk@uleth.ca, lucas@uleth.ca, christina.nord@uleth.ca, steph.varsanyi@uleth.ca, tyler.bonnell@uleth.ca
https://banzilab.github.io/

doi:10.1017/S0140525X21000558, e4

Abstract

Yarkoni argues that one solution is to abandon quantitative methods for qualitative ones. While we agree that qualitative methods are undervalued, we argue that both are necessary for thoroughgoing psychological research, complementing one another through the use of causal analysis. We illustrate how directed acyclic graphs can bridge qualitative and quantitative methods, thereby fostering understanding between different psychological methodologies.

Yarkoni stresses that a mismatch between intended verbal theories and hypotheses and current quantitative methods is a primary driver behind the generalizability crisis in psychology. Yarkoni suggests a number of ways forward, and we, as early career researchers, would like to present an additional option that avoids what often appears as a choice between doing *either* qualitative or quantitative work. In fact, recent calls for causal analysis within statistical practice require a strong integration between qualitative and quantitative work (Gelman & Vehtari, 2020; Pearl & Mackenzie, 2018). Here, we argue that the increasing need for clear qualitative understanding in order to conduct good quantitative modeling can act as a bridge between methods traditionally, and falsely, portrayed as separate.

Using statistical models to make causal inferences presents many difficulties. However, current theoretical and analytical developments have made it possible to make causal assumptions explicit, testable, and interpretable (Pearl & Mackenzie, 2018; Textor, van der Zander, Gilthorpe, Liśkiewicz, & Ellison, 2017). By adopting causal analysis, researchers are forced to make their theoretical assumptions clear, including the relevant variables to the system under study, how such variables are influenced by one another, and how they interact. As such, researchers need sufficient qualitative understanding of their study system in order to develop appropriate quantitative models in order to avoid statistical confounds (McElreath, 2020).

Thus, the growing use of causal frameworks within statistical practice presents a promising means to reduce the mismatch between intended verbal and statistical hypotheses, as well as facilitate communication between competing schools of thought. For example, personality research in the 1930s illustrates the emergence of the conflict between using quantitative and qualitative methods, the sources of which are made apparent through the use of a causal analysis. Certain assumptions, such as that only the individual and test instrument are relevant to personality research, rendered the environment unimportant in determining personality, which cemented the American quantitative approach to psychological research (Vernon, 1933). For example, tests of submissive behavior were assumed to reveal stable aspects of behavior in all contexts, regardless of the order of the items presented, past experience, comprehension skills, etc.

In contrast, Kurt Lewin's Berlin group held different philosophical assumptions, most notably, that the experimental environment – including communication between the researcher and participant – is a necessary part of psychological investigations. Therefore, Lewin's group rejected the idea that psychological phenomena could be meaningfully studied in terms of test scores and took a more qualitative approach (Van der Veer, 2000). For example, Tamara Dembo, a student of Lewin's, considered the role of the environment, as well as investigator, on behavior. Such aspects were either ignored or actively rejected by American psychologists at the time, so much so that when Lewin migrated to America, his ideas were marginalized. Today, most of Lewin's early papers remain untranslated, and his qualitative methods are "excluded from the experimental mainstream of American psychology" (Danziger, 1994, p. 178).

The important take away from this dispute is that it was a result of unchecked assumptions, many of which, as noted in Yarkoni's argument, remain in use today. It is our view that such conflicts have hope of resolution through the use of a causal analysis. Causal analysis provides a method, and in many ways a language, for researchers to collaboratively compare predictions and findings.

For example, below we use a form of causal modeling, directed acyclic graphs, to illustrate the different assumptions of two hypothetical theories of behavior (Textor et al., 2017). Every directed acyclic graph generates a set of testable implied conditional independencies that describe the model. Conditional independencies allow researchers to check that their statistical modeling matches their theoretical understanding of relationships between the variables under consideration, and to identify statistical confounds.

The first theory is represented by the directed acyclic graph in Figure 1A and is similar to the assumptions of early German personality researchers. The second, Figure 1B, is similar to early (and current) American assumptions about personality. The central difference between these two hypothetical approaches is the role of environment; in A, environment is a common cause of personality and behavior, and would be necessary in any statistical model measuring the direct effect of personality on behavior. In B, environment is conditional on behavior only, and statistical models using this framework could include personality, environment, or both.

However, implementing a causal approach within psychology's current framework requires considerable changes, both in research practices and pedagogy. As early career researchers, we are distinctly familiar with how statistics is currently being taught at universities and how the current system emphasizes quantitative statistics, often with not much more than a cursory glance at what qualitative methods may offer. A shift in pedagogy is essential. Quantitative and qualitative methods need to be placed on equal footing if we hope to build on and improve current research. We need to emphasize both the value of constructing strong theoretical frameworks for our systems of interest, and adopt more focused teaching of qualitative methods, starting at the undergraduate level. In doing so, psychological research can

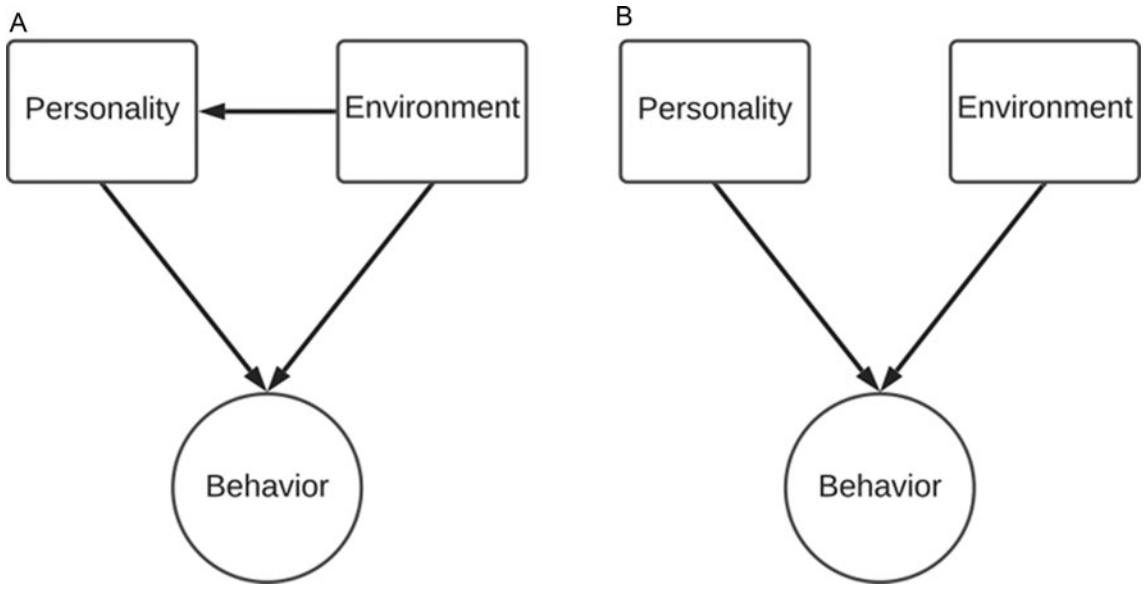

Figure 1. (Blersch et al.) Directed acyclic graphs representing two hypothetical theories of behavior. Arrows indicate the direction of the assumed causal relationships.

adopt a more structured mixed-methods approach already shown to be successful, and preferred, in other fields.

Mixed-methods approaches are increasingly popular (Bryman, 2008; Johnson & Onwuegbuzie, 2004), and we propose that directed acyclic graphs can serve as a bridge between qualitative and quantitative methods, overcoming this unnecessary methodological dichotomy while also facilitating well-informed quantitative analyses and communication between disciplines. That is, qualitative methods can help us construct directed acyclic graphs by identifying variables of interest, and then tested by subsequent quantitative statistics. Put more explicitly, we suggest quantitative researchers begin their study with a directed acyclic graph, and qualitative researchers end with one. By summarizing the study system using a causal analysis, researchers can easily identify what should be included (or might be missing) in subsequent research on the same topic or system.

Acknowledgments. Many thanks to the members of the Banzi Lab reading and writing group for inspiration.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Bryman, A. (2008). Bryman, Alan. Why do researchers integrate/combine/mesh/blend/mix/merge/fuse quantitative and qualitative research. In M. M. Bergman (Ed.), *Advances in mixed methods research* (pp. 86–100). Sage.

Danziger, K. (1994). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press.

Gelman, A., & Vehtari, A. (2020). What are the most important statistical ideas of the past 50 years. arXiv, 2012.00174v3. http://arxiv.org/abs/2012.00174v3.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26. https://journals.sagepub.com/doi/abs/10.3102/0013189x033007014.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. H. (2017). Robust causal inference using directed acyclic graphs: The R package "directed acyclic graphitty." *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341.

Van der Veer, R. (2000). Tamara Dembo's European years: Working with Lewin and Buytendijk. *Journal of the History of the Behavioral Sciences*, *36*(2), 109–126. https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1520-6696(200021)36:2 < 109::AID-JHBS1>3.0.CO;2-R.

Vernon, P. E. (1933). The American v. the German methods of approach to the study of temperament and personality. *British Journal of Psychology*, *24*(2), 156. https://search.proquest.com/openview/3d33b16cb13bc11e83b12c00f32def66/1?pq-origsite=gscholar&cbl=1818401.

Increasing generalizability via the principle of minimum description length

Wes Bonifay

Missouri Prevention Science Institute, University of Missouri, Columbia, MO 65211, USA
bonifayw@missouri.edu
https://education.missouri.edu/person/wes-bonifay/

doi:10.1017/S0140525X21000467, e5

Abstract

Traditional statistical model evaluation typically relies on goodness-of-fit testing and quantifying model complexity by counting parameters. Both of these practices may result in overfitting and have thereby contributed to the generalizability crisis. The information-theoretic principle of minimum description length addresses both of these concerns by filtering noise from the observed data and consequently increasing generalizability to unseen data.

As a remedy to the generalizability crisis, Yarkoni urges researchers to consider "cross-validation techniques that can minimize

overfitting and provide alternative ways of assessing generalizability outside of the traditional inferential statistical framework" (Sec. 3.6.7). I believe this advice is valuable and worthy of elaboration.

Traditional model evaluation techniques are beset by (at least) two inconvenient truths. First, goodness-of-fit (GOF) and generalizability are inextricably tied to model complexity (defined by Myung, Pitt, and Kim [2004] as "a model's inherent flexibility that enables it to fit a wide range of data patterns" [p. 12]). As models become more complex, GOF to the observed data increases, but generalizability to unseen data decreases. Additionally, GOF indices conflate fit to the useful signal in the data with fit to the useless noise, and so must be adjusted to account for complexity. The widely used Akaike Information Criterion (Akaike, 1973), for example, mitigates the effects of complexity by penalizing for the number of parameters.

However, this leads to the second issue: Complexity cannot be fully assessed by simply counting parameters (and in fact, overfitting can occur with just one parameter; Piantadosi, 2018). Complexity is also affected by the configuration of variables in the model (Cutting, Bruno, Brady, & Moore, 1992): Models that organize the same number of parameters in different configurations may differ in terms of GOF. It follows from these two issues that researchers who rely exclusively on GOF and quantify complexity only by counting parameters are exacerbating the generalizability crisis.

A solution to these problems can be found by bypassing probability theory altogether and adopting a technique from information theory. The principle of *minimum description length* (MDL; Rissanen, 1978, 1989) aims to separate regularity (i.e., meaningful information) from noise in the observed data and "squeeze out as much regularity as possible" (Grunwald, 2005, p. 15) via data compression. Suppose we have a sequence of nine binary digits that contains a regularity: twice as many 1s as 0s. The complete data space S includes $2^9 = 512$ patterns, but the regularity only applies to 84 (or 16.4%) of those patterns. Thus, our sequence belongs to a relatively small subset of S. A description (e.g., programming code) that compresses the complete data in this manner would be quite useful: We would know, for example, that future use of that code would return only those sequences that contain the same regularity.

According to the MDL principle, the best description (or model) is that which maximizes compression of S. Our nine-digit sequence could be further compressed: The regularity of "twice as many 1s as 0s + the first three digits are 1s" describes just 20 patterns, compressing the data to less than 4% of S. That is, over 96% of sequences would not follow this more precise regularity, so we should be "impressed" (in the sense of Meehl's [1990] rainfall analogy or Lakatos's [1978] example of Halley's comet) when we find a sequence that does.

What does this have to do with the generalizability crisis? In his introduction to MDL, Grunwald (2005) described two relevant features. First, "MDL procedures automatically and inherently protect against overfitting" (p. 5). GOF statistics may overfit the data by capturing both signal and noise, whereas MDL methods filter out that noise through data compression, allowing researchers to focus only on the signal. Second, "MDL methods can be interpreted as searching for a model with good predictive performance on unseen data" (p. 6). Mathematical proof of this statement can be found in Vitányi and Li (2000), who concluded that "compression of descriptions almost always gives optimal prediction" (p. 448).

Although MDL may seem obscure, consider it in light of this statement from Roberts and Pashler (2000) in *Psychological Review*, following their declaration that good fit cannot clarify what a theory predicts: "Without knowing how much a theory constrains possible outcomes, you cannot know how impressed to be when observation and theory are consistent" (p. 359). The phrase "a theory [that] constrains possible outcomes" can be rewritten in MDL terms as "a description that compresses the complete data space." Through that translation, it becomes clear that the MDL principle encapsulates Meehl's (1997) argument that "the narrower the tolerated range of observable values, the riskier the test, and if the test is passed, the stronger the corroboration of the substantive theory" (p. 407).

Various methods have been developed to quantify the MDL principle (see Myung, Navarro, & Pitt, 2006; Navarro, 2004; Pitt, Myung, & Zhang, 2002), but their formulations involve statistical obstacles such as integration across the complete data space. To sidestep this intractability, quantitative psychologists have relied on simulation methods to gain MDL-type insights regarding latent variable models. Preacher (2006) generated 10,000 random correlation matrices to simulate the complete continuous data space and fit competing structural equation models with the same number of parameters but different configurations to each matrix (interested readers can conduct similar MDL-type studies using the ockhamSEM package in R; Falk & Muthukrishna, 2021). Despite the fact that the number of parameters was held constant, certain models had an inherent tendency to fit better than others (termed "fitting propensity").

Bonifay and Cai (2017) expanded upon this work by considering the fitting propensity of several categorical data models. Among other findings, their analysis revealed that the confirmatory bifactor model achieved good fit to an excessively wide range of random datasets. The model was so deficient at compressing the data space (i.e., filtering out noise) that it accommodated an extremely wide range of data patterns, including many that were nonsensical. This MDL-inspired work demonstrated that good fit is essentially built into the bifactor model, so if the goal is to ensure generalizability, GOF testing should not be considered risky or severe (Watts, Poore, & Waldman, 2019).

In summary, the information-theoretic principle of MDL offers insights into overfitting and generalizability that are not possible using traditional methods. Although this principle may not address many of the generalizability issues described in the target article, it should be considered by researchers who wish to avoid overfitting and thereby enhance predictive accuracy.

Financial support. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210032.

Conflict of interest. None.

Note. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, *121*(3), 364–381.

Falk, C. F., & Muthukrishna, M. (2021). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000422.

Grünwald, P. (2005). A tutorial introduction to the minimum description length principle. In P. Grünwald, I.J. Myung, and M. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 3–81). MIT Press.
Lakatos, I. (1978). Introduction: Science and pseudoscience. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs* (pp. 1–8). Cambridge, UK: Cambridge University Press.
Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Erlbaum.
Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.
Myung, I. J., Pitt, M. A., & Kim, W. (2004). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *The handbook of cognition* (pp. 422–436). Thousand Oaks, CA: Sage.
Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*(2), 167–179.
Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation*, *16*(9), 1763–1768.
Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, *8*(9), 095118.
Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.
Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*(3), 227–259.
Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, *14*, 465–471.
Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific Publishing.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.
Vitányi, P. M., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446–464.
Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*(6), 1285–1303.

We need to be braver about the generalizability crisis

Todd S. Braver[a] and Sanford L. Braver[b]

[a]Department of Psychological & Brain Sciences, Washington University, St. Louis, St. Louis, MO 63130, USA and [b]Department of Psychology, Arizona State University, Tempe, AZ 85287, USA
tbraver@wustl.edu; sanford.braver@asu.edu
http://ccpweb.wustl.edu; http://www.public.asu.edu/~devra1/

doi:10.1017/S0140525X21000510, e6

Abstract

We applaud the effort to draw attention to generalizability concerns in twenty-first-century psychological research. Yet we do not feel that a pessimistic perspective is warranted. We outline a continuum of available methodological tools and perspectives, including incremental steps and meta-analytic approaches that can be readily and easily deployed by researchers to advance generalizability claims in a forward-looking manner.

We heartily applaud and commend Yarkoni for drawing attention to issues of generalizability in twenty-first-century psychological science. We strongly concur that these issues are of crucial importance; nevertheless, at least for most cognitive scientists and neuroscientists, they have not been prominent (in contrast to the voluminous literature in field and applied research, under the heading of *external validity*; e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979). Moreover, we agree with Yarkoni's perspective that the current preoccupation with reproducibility and replication may be misplaced, given that generalizability is a logically prior and potentially stronger concern. However, we definitely do not share Yarkoni's pessimistic perspective. Certainly, we do not support the extrapolation from this perspective to suggestions that academic psychologists consider pursuing different careers or switching from quantitative to qualitative research. Instead, we contend that there are ripe opportunities for psychological researchers to advance the generalizability of key phenomena of interest, by making greater use of the full continuum of available methods that can be deployed for this purpose.

The *radical randomization (RR)* experiment (Baribault et al., 2018; highlighted in Yarkoni) anchors one pole of this continuum, as the most ambitious and comprehensive strategy. An RR experiment involves *many* (16 in Baribault et al., 2018) potentially irrelevant factors – or moderators – that are varied *randomly* within the experimental design (as "micro-experiments"). With Bayesian hierarchical modeling, both the summary effect size and the moderating effect of each random factor can be properly estimated. However, as a highly effort- and resource-intensive endeavor, the RR experiment seems less likely to serve as the primary approach for addressing generalizability concerns.

Fortunately, more easily deployed approaches are available. We agree with Shadish, Cook, and Campbell (2002), who eschew both of the key defining features (italicized above) of RR studies: (1) that researchers simultaneously address multiple potential (often theoretically irrelevant) moderators at once in the same meta-study; and (2) that the levels of these moderating factors be both randomly selected and analyzed as random- (rather than fixed-) factors. The objection to (1) is based on the insight that there will always remain a virtually infinite space of additional, possible, non-varied factors that could limit generalizability. Indeed, the very nature of inductive logic precludes ever completely resolving generalizability issues. Nevertheless, some inferential purchase can be provided – albeit somewhat more slowly – via an *incremental* (i.e., study-by-study), rather than comprehensive, strategy. With respect to (2), although random selection and random-effects models are clearly preferred, researchers can still legitimately advance the generalizability of their postulates by "guessing at laws and checking out some of these generalizations in other *equally specific* but *different* conditions" (Campbell & Stanley, 1963, p. 17, emphasis added).

Thus, to anchor the other pole of generalizability efforts, we propose that researchers consider varying at least one, unique, and supposedly irrelevant contextual factor in each experiment (see Yarkoni, p.9 for examples). Importantly, even with only a few (but of course more than one) *purposively* selected levels of this factor, there is still an interpretational advantage to be gained. Specifically, even if using fixed-effects rather than random-effects analysis, the *interaction* of this factor with the main effect of interest can be tested, to estimate its impact. Only if the interaction effect is small and insignificant can the claim be made that the induced heterogeneity is indeed plausibly irrelevant; if so, generalizability claims over this factor can be furthered. For example, imagine if, in the original Schooler and Engstler-Schooler (1990) study highlighted by Yarkoni, multiple perpetrator videos had been used, with similar effect sizes for each (i.e., no interaction). Moreover, this approach enables generalizability claims regarding a phenomenon of interest to be advanced incrementally, study-by-study. Generalizability claims become more grounded and justifiable – albeit with greater effort – by moving

along the continuum toward RR: varying additional putative nuisance factors, including more exemplars of each factor, selecting (sampling) these exemplars at random rather than purposively, and evaluating them with random-effects rather than fixed-effects analyses.

The arguably mid-continuum issue of conceptual, as opposed to exact, replications also highlights our key disagreement with Yarkoni. In particular, we claim that Yarkoni unfairly undersells the substantial epistemological advantage of conceptual replications. Of course, he is not alone: researchers often implement replications by trying to precisely match all the details of an initial study, presumably out of a superstitious desire to "get it right." Yet, as many have long pointed out (Brunswik, 1956; Campbell & Stanley, 1963; Cronbach, 1975), the stronger alternative is to purposely vary the features that should be theoretically irrelevant, with the goal of finding that such variation does not in fact alter the outcome. Yarkoni dismisses conceptual replications, by alleging that they "do not lend themselves well to a coherent modeling strategy.... It is rarely obvious how one can combine the results to obtain a meaningful estimate of the robustness or generalizability of the common effect (p.25)."

We strongly disagree with Yarkoni on this critical point. In particular, meta-analysis techniques are precisely designed to evaluate the robustness of effect size findings over a set of studies. In addition to statistics that quantify summary (overall) effect size, it is standard to also evaluate homogeneity of effect, i.e., through indices as the Q-test and the I^2 statistics (Hedges & Olkin, 1985). Meta-analysis is typically invoked for retrospective reviews of a body of literature, yet Braver, Thoemmes, and Rosenthal (2014) extend its utility via the *continuously cumulating meta-analytic* approach. With this approach, meta-analytic calculations can be employed incrementally – even within-study (i.e., across experiments) – as new findings emerge. Newer Bayes Factor approaches may gain even greater traction as a means of directly implementing this incremental perspective (Scheibehenne, Jamil, & Wagenmakers, 2016).

In summary, our goal is to help psychological researchers appreciate that there is an entire buffet of experimental design and analysis options readily available and waiting to be deployed to address the issues of generalizability. There is no need to despair, or begin searching out alternative career choices! Indeed, all that is needed is for the field to face the generalizability crisis in a *Braver* manner.

Financial support. TSB acknowledges the following funding sources, which supported this work: R37 MH066078, R21 AG067296, T32 NS115672, NSF NCS-FO 1835209.

Conflict of interest. None.

References

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 201708285. https://doi.org/10.1073/pnas.1708285114.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand-McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand-McNally.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/https://psycnet-apa-org.libproxy.wustl.edu/doi/10.1037/h0076829.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press. https://doi.org/https://doi.org/10.1016/C2009-0-03396-0.

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results. *Psychological Science*, *27*(7), 1043–1046. https://doi.org/10.1177/0956797616644081.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71. https://doi.org/10.1016/0010-0285(90)90003-m.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin and Company.

Impact on the legal system of the generalizability crisis in psychology

Chris R. Brewin

Department of Clinical, Educational & Heatlh Psychology, University College London, London WC1E 6BT, UK
c.brewin@ucl.ac.uk

doi:10.1017/S0140525X21000480, e7

Abstract

Overgeneralizations by psychologists of the research evidence on memory and eyewitness testimony, such as "memory decays with time" or "memories are fluid and malleable," are beginning to appear in legal judgements and guidance documents, accompanied by unwarranted disparagement of lay beliefs about memory. These overgeneralizations could have significant adverse consequences for the conduct of civil and criminal law.

The generalizability crisis so ably articulated by Yarkoni can lead to particularly unfortunate consequences in applied fields such as the law. Some of the most egregious examples of this occur in the fields of memory and eyewitness testimony. General statements are made asserting that lay opinions about memory, including those of jurors and lawyers, are frequently in error (Berkowitz & Frenda, 2018; Clifasefi, Garry, & Loftus, 2007; Lynn, Evans, Laurence, & Lilienfeld, 2015). Memory is claimed to be error-prone or unreliable without the qualification that it may be accurate under other conditions (Lynn & Payne, 1997; Zajac, Garry, London, Goodyear-Smith, & Hayne, 2013). Some psychologists refer to "laws" of memory (Howe, 2013), even though memory phenomena are known to be highly dependent on such factors as who the participants are, the conditions present at encoding, what is being recalled, the encoding conditions, and how memory is assessed (Roediger, 2008). Such overgeneralizations can then become incorporated in legal and judicial documents.

The general public is supposed to think that memory involves playing back events exactly as they happened, literally "like a video camera" (Lacy & Stark, 2013; Lilienfeld, Lynn, Ruscio, & Beyerstein, 2010). This often repeated, but false, assertion involves generalization from responses to single survey questions such as "Human memory works like a video camera, accurately recording the events we see and hear so that we can review and inspect them later" (Simons & Chabris, 2011). In fact most people do experience some memories, particularly those involving important

events, as a connected series of scenes rather like a videotape. Simply asking additional questions reveals that memory beliefs are much more nuanced than this and people are well aware that their recollection is not always reliable (Brewin, Li, Ntarantana, Unsworth, & McNeilis, 2019). However, the idea that the public had mistaken ideas about memory led to the state of New Jersey instructing jurors that "Research has revealed that human memory is not like a video recording that a witness need only replay to remember what happened. Human memory is far more complex" (New Jersey Courts, 2020). The danger here is that testimony could be discounted purely on the basis that the witness described their memory as "like a video recording," when this statement did not at all imply a naïve or mistaken view of memory.

For many years eyewitness confidence was thought to be only weakly related to accuracy, but this conclusion was overturned when techniques for analyzing the question were improved (Wixted & Wells, 2017). Eyewitness confidence is not a guarantee of accuracy, but is very highly related to accuracy when memory is uncontaminated and suitable interviewing procedures are used (Wixted, Mickes, & Fisher, 2018). However, the inappropriate generalization had already been incorporated into compulsory jury guidance issued by the Supreme Court of New Jersey (http://www.judiciary.state.nj.us/pressrel/2012/pr120719a.htm) for cases involving eyewitness identification, which stated "Although some research has found that highly confident witnesses are more likely to make accurate identifications, eyewitness confidence is generally an unreliable indicator of accuracy" (p. 4).

Witnesses have often been impugned by claims that memory is unreliable, for example, invoking research on "false memory" (Brewin, Andrews, & Mickes, 2020; Wade, Nash, & Lindsay, 2018; Wixted et al., 2018). This term is applied to several different experimental paradigms such as associative illusions (the Deese-Roediger-McDermott paradigm), manipulating memory for detail by providing misleading post-event information, artificially inflating the perceived likelihood that a non-remembered event occurred, and implanting memories for childhood events that never happened. Implications for the legal system are then discussed without distinguishing between these very different paradigms, and without recognition that there is no overall proclivity to experience all types of false memory (Lacy & Stark, 2013). The fact that memory can be manipulated in the laboratory is important but does not by itself allow of any conclusion about how reliable memory is under normal, real life circumstances. Despite this, some organizations have endorsed statements such as "Science shows that the memory of an honest witness who gives evidence in international arbitration proceedings can easily become distorted" (International Chamber of Commerce, 2020).

Similarly, it is sometimes suggested on the basis of memory implantation research that it is relatively simple to create false memories (Conway, 2012). This overgeneralization overlooks the special procedures that are used, including at times a high degree of deception, and the fact that only a small minority of participants may succumb to them. Little is known about the durability of the effects or whether they are associated with the degree of conviction necessary to sustain legal procedures such as cross-examination (Brewin & Andrews, 2017).

Overgeneralizations also occur when defence experts comment on contemporary witnesses reporting historic crimes. The simple idea that "memory decays with time" is sometimes put forward, without acknowledgment that most studies involved meaningless experimental materials and that there are numerous counter-examples (Roediger, 2008). Moreover, the claim does not make clear that with significant personal experiences initial decay in the total amount recalled typically plateaus, resulting in stable and largely accurate long-term recall of the remainder (Diamond, Armson, & Levine, 2020; Hirst et al., 2015). Despite this, the compulsory jury guidance referenced above contained the generalization: "Memories fade with time…. In other words, the more time that passes, the greater the possibility that a witness's memory of a perpetrator will weaken" (p. 5).

In 2013, a UK judge, Justice Leggatt, commented "Psychological research has demonstrated that memories are fluid and malleable, being constantly rewritten whenever they are retrieved (EWHC 3560 (Comm); Case No. 2011 Folio 1267). A person would be in error to suppose: that the stronger and more vivid is our feeling or experience of recollection, the more likely the recollection is to be accurate." These statements do not accurately reflect a complex literature, and fail to take into account the number of times an event happens, its importance, and memory rehearsal, whether deliberate or spontaneous. Legal professionals are not in a position to know that these are overgeneralizations. Psychologists, as Yarkoni has demonstrated, are the ones who are responsible and this can have serious consequences when their statements are repeated in the real world.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Berkowitz, S. R., & Frenda, S. J. (2018). Rethinking the confident eyewitness: A reply to Wixted, Mickes, and Fisher. *Perspectives on Psychological Science*, *13*(3), 336–338. https://doi.org/10.1177/1745691617751883

Brewin, C. R., & Andrews, B. (2017). Creating memories for false autobiographical events in childhood: A systematic review. *Applied Cognitive Psychology*, *31*, 2–23. https://doi.org/10.1002/acp.3220

Brewin, C. R., Andrews, B., & Mickes, L. (2020). Regaining consensus on the reliability of memory. *Current Directions in Psychological Science*, *29*(2), 121–125, Article 0963721419898122. https://doi.org/10.1177/0963721419898122

Brewin, C. R., Li, H., Ntarantana, V., Unsworth, C., & McNeilis, J. (2019). Is the public understanding of memory prone to widespread "myths"? *Journal of Experimental Psychology: General*, *148*, 2245–2257. https://doi.org/10.1037/xge0000610

Clifasefi, S. L., Garry, M., & Loftus, E. F. (2007). Setting the record (or video camera) straight on memory: The video camera model of memory and other memory myths. In S. Della Sala (Ed.), *Tall tales about the mind and brain: Separating fact from fiction* (pp. 60–75). Oxford University Press.

Conway, M. A. (2012). Ten things the law and others should know about human memory. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 359–371). Oxford University Press.

Diamond, N. B., Armson, M. J., & Levine, B. (2020). The truth is out there: Accuracy in recall of verifiable real-world events. *Psychological Science*, *31*(12), 1544–1556. https://doi.org/10.1177/0956797620954812

Hirst, W., Phelps, E. A., Meksin, R., Vaidya, C. J., Johnson, M. K., Mitchell, K. J., … Olsson, A. (2015). A ten-year follow-up of a study of memory for the attack of September 11, 2001: Flashbulb memories and memories for flashbulb events. *Journal of Experimental Psychology-General*, *144*(3), 604–623. https://doi.org/10.1037/xge0000055

Howe, M. L. (2013). Memory lessons from the courtroom: Reflections on being a memory expert on the witness stand. *Memory (Hove, England)*, *21*(5), 576–583. https://doi.org/10.1080/09658211.2012.725735

International Chamber of Commerce. (2020). *The accuracy of fact witness memory in international arbitration*. Retrieved from https://iccwbo.org/content/uploads/sites/3/2020/11/icc-arbitration-adr-commission-report-on-accuracy-fact-witness-memory-international-arbitration-english-version.pdf

Lacy, J. W., & Stark, C. E. L. (2013). The neuroscience of memory: Implications for the courtroom. *Nature Reviews Neuroscience*, *14*(9), 649–658. https://doi.org/10.1038/nrn3563

Lilienfeld, S. O., Lynn, S. J., Ruscio, J., & Beyerstein, B. L. (2010). 50 *Great myths of popular psychology: Shattering widespread misconceptions about human behavior*. Wiley-Blackwell.

Lynn, S. J., Evans, J., Laurence, J. R., & Lilienfeld, S. O. (2015). What do people believe about memory? Implications for the science and pseudoscience of clinical practice. *Canadian Journal of Psychiatry*, *60*(12), 541–547. https://doi.org/10.1177/070674371506001204

Lynn, S. J., & Payne, D. G. (1997). Memory as the theater of the past: The psychology of false memories. *Current Directions in Psychological Science*, *6*(3), 55–55. https://doi.org/10.1111/1467-8721.ep11512651

New Jersey Courts (2020). Model Criminal Jury Charges. Identification: In-court and out-of-court identifications. Retrieved 20 December 2021 from https://www.njcourts.gov/attorneys/assets/criminalcharges/idinout.pdf

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254. https://doi.org/10.1146/annurev.psych.57.102904.190139

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PLoS ONE*, *6*. https://doi.org/10.1371/journal.pone.0022757

Wade, K. A., Nash, R. A., & Lindsay, D. S. (2018). Reasons to doubt the reliability of eyewitness memory: Commentary on Wixted, Mickes, and Fisher (2018). *Perspectives on Psychological Science*, *13*(3), 339–342. https://doi.org/10.1177/1745691618758261

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Zajac, R., Garry, M., London, K., Goodyear-Smith, F., & Hayne, H. (2013). Misconceptions about childhood sexual abuse and child witnesses: Implications for psychological experts in the courtroom. *Memory*, *21*(5), 608–617. https://doi.org/10.1080/09658211.2013.778287

Exposing and overcoming the fixed-effect fallacy through crowd science

Wilson Cyrus-Lai[a], Warren Tierney[b], Martin Schweinsberg[c] and Eric Luis Uhlmann[a]

[a]Organisational Behaviour Area, INSEAD, Singapore; [b]Organisational Behaviour Area/Marketing Area, INSEAD, Singapore and [c]Martin Schweinsberg, Organisational Behaviour Area, ESMT Berlin, 10178, Berlin Germany.
wilson-cyrus.lai@insead.edu; eric.luis.uhlmann@gmail.com
warren.tierney@insead.edu
martin.schweinsberg@esmt.org

doi:10.1017/S0140525X21000297, e8

Abstract

By organizing crowds of scientists to independently tackle the same research questions, we can collectively overcome the generalizability crisis. Strategies to draw inferences from a heterogeneous set of research approaches include *aggregation*, for instance, meta-analyzing the effect sizes obtained by different investigators, and *parsing*, attempting to identify theoretically meaningful moderators that explain the variability in results.

Yarkoni highlights the *fixed-effect fallacy*, arguing that many if not most research findings are unlikely to prove robust to stimulus sampling and task operationalizations. Experimental studies in psychology and related fields are exposed to the possibility that the effect is specific to the stimulus set in question, such that alternative approaches could have attenuated or even reversed the reported finding. Recent initiatives to crowdsource the analyses of complex datasets (Bastiaansen et al., 2020; Botvinik-Nezer et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018), and the design of experiments (Baribault et al., 2018; Landy et al., 2020) provide strong quantitative evidence for these assertions. When different scientists independently analyze the same dataset to try and answer the same research question, or separately create their own experimental design to test the same hypothesis, a wide range of results are obtained.

These large-scale crowd science projects illustrate two key approaches to drawing robust conclusions and building strong theory through diversity in approaches and results. One strategy to overcoming the generalizability challenge is *aggregation*, for example, simply meta-analyzing across the estimates obtained by independent analysts or from different experimental designs. Another is *parsing*, or attempting to find meaningful moderators that explain why some approaches yield large estimates and others small to null estimates or even estimates reversed in sign.

The parsing strategy is in harmony with the perspectivist approach to theoretical progress, which assumes that most phenomena in the social sciences are massively moderated (McGuire, 1973, 1983). From this perspective, "the opposite of a great truth is also true" (Banaji, 2003), and thus it is unsurprising that different empirical approaches to testing the same idea can return effect size estimates that are opposed in sign. The fundamental task of researchers, from a perspectivist standpoint, is to untangle this web by identifying moderators that will allow us to predict when effects emerge, disappear, and reverse. However, we suggest that aggregation and parsing can be complementary rather than competing: meta-scientists can both meta-analyze across crowdsourced approaches and seek to meaningfully explain variability in effect sizes.

In an illustration of the aggregation strategy, Landy et al. (2020) recruited up to 13 research teams to independently create experimental stimulus sets to test the same set of five original hypotheses, all supported in unpublished research by the original authors (e.g., "working for no reason is morally praised," "deontologists are happier than consequentialists"). Over 15,000 research participants were randomly assigned to the different study designs. All five original effects directly replicated using the same stimulus set the original authors had used. However, four of five hypotheses had different material – makers created designs that returned statistically significant effects in opposite directions from one another. At the same time, two out of five original hypotheses proved conceptually robust when meta-analyzing the results across the experimental designs from the different teams of researchers. This maps on closely to predictions by Yarkoni and others, that even when directly replicable, only a minority of findings in social psychology and related fields will prove generalizable across contexts and approaches.

Employing both the aggregation and parsing strategies together, Schweinsberg et al. (2021) asked up to 15 independent researchers to test two hypotheses using the same dataset capturing gender and status dynamics in intellectual debates. Not only statistical choices (e.g., covariates), but also the operationalization of variables (e.g., status) were left unconstrained and up to the individual researchers' discretion. For example, an analyst could choose to identify high versus low status academics using job rank, citation counts, PhD institution rank, or a combination of indicators. No two researchers employed the same specification. For both hypotheses, independent analysts reported statistically significant estimates in opposite directions despite relying on the same dataset. Hypothesis 1 (women speak more in the presence of other women) was supported while aggregating across different measurement and testing approaches, whereas Hypothesis 2 (high status academics speak more) was comparatively not, with estimates distributed around zero in the latter case. Leveraging a Boba multiverse analysis (Liu, Kale, Althoff, &

Heer, 2020; see also Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) to identify key analyst choice points, Schweinsberg et al. (2021) further demonstrate that differing variable operationalizations directly contributes to this radical dispersion in estimates across different analysts. For example, researchers who operationalized status as job rank consistently returned negative estimates for H2, whereas those operationalizing status using ranking of doctoral institution returned consistently positive estimates. This illustrates how the parsing strategy treats variability across different approaches as clues to meaningful moderation, rather than error to be averaged away.

In order to draw generalizable conclusions, Tierney et al. (in preparation) assigned teams of doctoral students and professors to separately create conceptual replication designs testing for backlash against angry women. The original study finds that although male managers who express anger (relative to sadness or neutral emotions) experience a boost in status, female managers who express anger are accorded less social status and respect (Brescoll & Uhlmann, 2008). Participants in this ongoing data collection across over 50 laboratories are randomly assigned to one of 27 study designs (the original design and 26 conceptual replication designs) testing the hypothesized interaction between target gender and emotion expression. The employed methods include scenarios, ostensive newspaper stories, audio recordings, video recordings, and storyboards with illustrated characters as well as a myriad of different ways of expressing anger. In addition to a preregistered meta-analysis of the results across designs, we will systematically test potential moderators of the results across designs; among these are anger extremity, dominance displays, and the salience of target gender.

In summary, we can collectively overcome the generalizability crisis by organizing crowds of scientists to tackle the same research questions independently. Doing so will further expose the fixed-effect fallacy that a single analysis and research paradigm are sufficient for drawing strong theoretical inferences. Scientists can rely on the wisdom of the crowd by aggregating results across independent investigators, and seek to identify meaningful moderators of the results across different approaches, in the perspectivist spirit.

Financial support. This research was supported by an R&D grant from INSEAD to Eric Uhlmann.

Conflict of interest. None.

References

Banaji, M. R. (2003). The opposite of a great truth is also true: Homage of Koan #7. In J. Jost, D. Prentice & M. R. Banaji (Eds.), *The yin and yang of progress in social psychology: Perspectivism at work* (pp. 127–140). Washington, DC: American Psychological Association.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., … Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M. … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*, 84–88.

Brescoll, V., & Uhlmann, E. L. (2008). Can angry women get ahead? Status conferral, gender, and workplace emotion expression. *Psychological Science*, *19*, 268–275.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479.

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2020). Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1753–1763.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, *26*(3), 446–456.

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovations and reform in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. *16*, pp. 1–47). New York, NY: Academic Press.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O., van Aert, R., van Assen, M., Liu, Y., … Uhlmann, E. (2021). Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organizational Behavior and Human Decision Processes*, *165*, 228–249.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. … Nosek, B. N. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Tierney, W., Cyrus-Lai, W., … Uhlmann, E. L. (in preparation). Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral. Crowdsourced research project in progress.

Separate substantive from statistical hypotheses and treat them differently

Mike Dacey

Department of Philosophy, Bates College, Lewiston, ME 04240, USA.
mdacey@bates.edu; mikedacey.net

doi:10.1017/S0140525X21000157, e9

Abstract

I suggest addressing the problems Yarkoni identifies by separating substantive from statistical hypotheses, and treating them differently. A statistical test of experimental data only bears directly on statistical hypotheses. Evaluation of related substantive hypotheses requires an additional, qualitative inference to the best explanation. Statistical inference cannot do all of the work of theory choice.

The target article highlights a vital problem in psychology: the inferential gap between statistical models and verbally-expressed psychological theories is too rarely appreciated or respected. However, I am perhaps more optimistic that the core problem can be solved.

My suggestion is, in short, to recognize the distinction in statistics between *statistical hypotheses* and *substantive hypotheses*, and to treat them differently from one another (e.g., Hays, 1994). In psychology, a statistical hypothesis simply describes the distribution of a trait across a population, such as a behavioral tendency or level of task performance. A substantive hypothesis makes claims about the causal structure responsible for that distribution, such as the function of the cognitive systems operating. Yarkoni's paraphrase of a conservative conclusion about ego depletion could describe a statistical hypothesis: "crossing out the letter *e* slightly decreases response accuracy on a subsequent Stroop task" (sect. 6.3.1, para. 2). The attending substantive hypothesis might be "psychological processes requiring attention are subject to ego depletion."

Only the statistical hypothesis is directly tested using statistical methods. The substantive hypothesis must be evaluated separately

(though not entirely independently). The result is a two-step process of evaluation (see Dacey, forthcoming).

Each step in the process implements one of Yarkoni's suggested courses of action. The first step is statistical inference *proper*, evaluating a statistical hypothesis based on the data by one's preferred statistical method. This step implements Yarkoni's suggestion that we draw more conservative inferences (sect. 6.3.1). The statistical result only bears directly on the limited statistical hypothesis, such as Yarkoni's limited conclusion, quoted above. The second step is the evaluation of the substantive hypothesis, which should implement Yarkoni's suggestion that we embrace qualitative analysis (sect. 6.2). This step requires evaluating how the decision about the statistical hypothesis bears on the substantive hypothesis, taking into account other relevant evidence. This is an inference to the best explanation that is not likely well-modeled by quantitative tools or formal logic (*contra* concerns about affirming the consequent; section 6.3.6).

Crucially, this means that a single statistical result will usually be very weak evidence for the substantive hypotheses of interest. The mind is complicated and, as Yarkoni highlights, there are many sources of variance, so it is very rare that one substantive hypothesis cannot explain, or at least accommodate, a statistical result. The statistics simply say "here is an effect that our theories should explain." We decide which candidate explanation is best in the second, qualitative step. This must consider all relevant findings as targets of explanation, not just the most recent: we should resist viewing a single experiment as a stand-alone test of competing substantive hypotheses. I'd even suggest that many experiments be seen as one part of a larger project of characterizing or mapping the capacities involved, not as tests of one substantive hypothesis against another at all. It is a mistake to try to make statistical inference do all of the work of theory choice.

This approach will not fix all of the problems the target article mentions, but it would go a long way towards addressing the core problem. I take this to require changes in the way findings are reported in empirical papers, and perhaps in the way theoretical interpretations are argued for in review papers. However, I don't take it to require any drastic change to the science.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Dacey, M. (forthcoming). Evidence in default: Rejecting default models of animal minds. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/714799.

Hays, W. L. (1994). *Statistics*. Harcourt Brace College Publishers.

Measurement practices exacerbate the generalizability crisis: Novel digital measures can help

Brittany I. Davidson[a,b], David A. Ellis[a], Clemens Stachl[c], Paul J. Taylor[d] and Adam N. Joinson[a]

[a]School of Management, University of Bath, Claverton Down, Bath BA2 7AY, UK; [b]Department of Engineering, University of Bristol, Bristol BS1 5DD, UK; [c]Institute of Behavioral Science & Technology, University of St. Gallen, CH-9000, Switzerland and [d]Department of Psychology, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK

bid23@bath.ac.uk
https://www.brittanydavidson.co.uk/
dae30@bath.ac.uk
http://www.davidaellis.co.uk/
cstachl@stanford.edu
https://www.clemensstachl.com
p.j.taylor@lancaster.ac.uk
https://pauljtaylor.com/
aj266@bath.ac.uk
http://www.joinson.com/home/Welcome.html

doi:10.1017/S0140525X21000534, e10

Abstract

Psychology's tendency to focus on confirmatory analyses before ensuring constructs are clearly defined and accurately measured is exacerbating the generalizability crisis. Our growing use of digital behaviors as predictors has revealed the fragility of subjective measures and the latent constructs they scaffold. However, new technologies can provide opportunities to improve conceptualizations, theories, and measurement practices.

Yarkoni highlights the disconnect between psychology's descriptive theories and its inferential tests – a problem we argue is exacerbated by inadequate measurement. The primacy of measurement in psychology's history has ebbed-and-flowed, from the absolute focus on what was observable and quantifiable that defined behaviorist approaches (Hayes & Brownstein, 1986; Skinner, 1963, 1976) to the overreliance on button presses and mouse clicks that characterizes some modern research (Baumeister, Vohs, & Funder, 2007). Today, digital trace data provide new opportunities for rich measurement that captures behavioral, situational, and environmental/contextual factors simultaneously (Lazer et al., 2020; Mischel, 2004). For instance, smartphones are a powerful data source – a collection of sensors and logging routines that we carry with us for large swathes of the day – that psychologists are utilizing to predict a variety of outcomes, from social interaction, personality, mood, to general health (Davidson, 2020; Ellis, 2020; Harari et al., 2020; Miller, 2012; Piwek, Ellis, Andrews, & Joinson, 2016; Stachl et al., 2020).

Improved methodology alone will not result in rapid progress for the behavioral sciences (see Kaplan, 1964; Uttal, 2001). For example, digital trace data have re-ignited problems with traditional operationalizations of latent variables. Research demonstrating associations between new and old measures often fails to articulate why a connection between a latent measure (e.g., mood disturbance) and a behavioral (digital) predictor (e.g., keystroke speed) should exist in advance of an analysis (Davidson, 2020; Zulueta et al., 2018). Without specification or theory, the focus on prediction over explanation restricts generalizability further. A related challenge is the disconnect between subjective and objective measures (e.g., Taylor et al., 2021), where predictive studies find their survey data predict an outcome, but objective measures do not (Eisenberg et al., 2019). Here, the problem is an overreliance on subjective methodologies to measure both latent and observable constructs. For example, the gold standard for personality measurement relies on surveys (e.g., HEXACO, OCEAN, Big 5) and remains contested (Cattell, 1958; Kagan, 2001). Similarly, other measures including estimates of everyday behavior rarely align with reality (Parry et al., 2020). While latent measurement remains core to psychological science, many

constructs are developed rapidly, with little standardization, and rely on face validity alone (e.g., "internet addiction," despite being sardonic in origin, has spawned 100s of technology addiction scales; Howard & Jayne, 2015). New digital sources need to avoid these issues if they are to prosper.

Illuminating the complex relationship between generalizability and measurement further – observations of behavior via digital traces will often only explain (or predict) part of a broad latent construct. At face value, predicting part of extraversion may appear straightforward from digital recordings of speech, or time spent using social apps. However, there are other subcomponents of extraversion that these data will struggle to explain (e.g., feeling indifferent to social activities). Other personality factors such as openness and agreeableness remain conceptually more challenging to map onto (a single) digital behavior (Hinds & Joinson, 2019; Stachl et al., 2020). Hence, it is critically important psychology shifts away from predictive validity alone as evidence for successful operationalization and parameterization, especially from new data sources (Boyd, Pasca, & Lanning, 2020). Any new digital measure has to be developed incrementally, where researchers first describe how it conceptually aligns with an existing latent construct (Glewwe & van der Gaag, 1990). Assuming that digital traces are behavioral expressions of latent variables, researchers should be able to qualitatively express links at a more general level first across contexts, then move to specifics, which would enhance generalizability.

Of course, refocusing on actual behavior via digital traces will not be a panacea. Some digital traces may be "objective," but they are rarely error-free (Sen, Floeck, Weller, Weiss, & Wagner, 2019). For example, a microphone-based audio classifier can detect whether ambient conversations are taking place around an individual, but it may not distinguish real conversations from someone watching television. Similarly, little consideration is given to how measurement variance might be reduced or maximized for a new digital source. For example, while some assessments in psychology (e.g., cognitive tasks) do not produce reliable individual differences, others (e.g., mood) purposefully reflect variations in individual responses (Hedge, Powell, & Sumner, 2018). Hence, it is critical to find ways to share *raw* data, processing pipelines, and analysis scripts for digital trace research, as the degrees of freedom are vast, which causes large variance in conclusions made from the same data (Silberzahn et al., 2018; Towse, Ellis, & Towse, 2020). Validation procedures are likely to reflect the disparity of digital data sources, but combining small and large-scale approaches (e.g., $N = 1$ sample, case studies) can successfully quantify errors associated with smartphone sensing-based methods (Geyer, Ellis, Shaw, & Davidson, 2020; Sen et al., 2019; Szot, Specht, Specht, & Dabrowski, 2019). Only then can related work explore how signals from multiple systems may be combined to improve data efficiency. Failure to ensure this basic research is completed will result in little progress as research agendas risk shifting in the wrong direction if the grounding principles are weak, particularly in applied settings, such as security and health, which are increasingly interested in digital traces (Davidson, 2020; Guttman & Greenbaum, 1998).

Moreover, we acknowledge that research in this space remains challenging to conduct because data derived from digital sources can be difficult to access, handle, and interpret (DeMasi, Kording, & Recht, 2017). This challenges the way psychologists are trained and incentivized (not) to publish descriptive findings in an interdisciplinary landscape. However, we are hopeful that new methods and emerging forms of data will complement psychology's diverse measurement practices. Collectively termed the *Internet of Things*, the future potential for data linkage that could further leverage real-world research remains an exciting prospect. In the long term, taking time to understand how behavioral, situational, and environmental/contextual factors can be extracted from objective digital data will allow psychology to develop robust contextualized and comprehensive theory (Lazer et al., 2020).

Our muse are people and psychology should critically consider how it moves forward and merges old and new. Generalizability requires sound measures first, but there is still little agreement between psychologists on what is worth measuring.

Financial support. This work was part-funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1 to PJT; ANJ; DAE), www.crestresearch.ac.uk and by the National Science Foundation (SES-1758835 to CS).

Conflict of interest. None.

References

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x.

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through Big behavioural data. *European Journal of Personality*, *34*(5), 599–612. https://doi.org/10.1002/per.2254.

Cattell, R. B. (1958). What is "objective" in "objective personality tests"? *Journal of Counseling Psychology*, *5*(4), 285. https://doi.org/10.1037/h0046268.

Davidson, B. I. (2020). The crossroads of digital phenotyping. *General Hospital Psychiatry*. https://doi.org/10.1016/j.genhosppsych.2020.11.009.

DeMasi, O., Kording, K., & Recht, B. (2017). Meaningless comparisons lead to false optimism in medical machine learning. *PLoS ONE*, *12*(9), e0184604. https://doi.org/10.1371/journal.pone.0184604.

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1.

Ellis, D. A. (2020). *Smartphones within psychological science*. Cambridge University Press.

Geyer, K., Ellis, D. A., Shaw, H., & Davidson, B. I. (2020). Open source smartphone app and tools for measuring, quantifying, and visualizing technology use [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/eqhfa.

Glewwe, P., & van der Gaag, J. (1990). Identifying the poor in developing countries: Do different definitions matter? *World Development*, *18*(6), 803–814. https://doi.org/10.1016/0305-750X(90)90003-G.

Guttman, R., & Greenbaum, C. W. (1998). Facet theory: It's development and current status. *European Psychologist*, *3*, 13–36. https://doi.org/10.1027/1016-9040.3.1.13.

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., … Gosling, S. D. (2020). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, *119*(1), 204–228. https://doi.org/10.1037/pspp0000245.

Hayes, S. C., & Brownstein, A. J. (1986). Mentalism, behavior-behavior relations, and a behavior-analytic view of the purposes of science. *The Behavior Analyst*, *9*(2), 175–190. https://doi.org/10.1007/BF03391944.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science*, *28*(2), 204–211.

Howard, M. C., & Jayne, B. S. (2015). An analysis of more than 1,400 articles, 900 scales, and 17 years of research: The state of scales in cyberpsychology, behavior, and social networking. *Cyberpsychology, Behavior, and Social Networking*, *18*(3), 181–187.

Kagan, J. (2001). The need for new constructs. *Psychological Inquiry*, *12*(2), 84–103. https://doi.org/10.1207/S15327965PLI1202_03.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. Chandler Publishing Company.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., … Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*(1), 1–22. https://doi.org/10.1146/annurev.psych.55.042902.130709.
Parry, D. A., Davidson, B. I., Sewall, C., Fisher, J. T., Mieczkowski, H., & Quintana, D. (2020). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. PsyArXiv. doi:10.31234/osf.io/f6xvz.
Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The rise of consumer health wearables: Promises and barriers. *PLoS Medicine*, *13*(2), e1001953. https://doi.org/10.1371/journal.pmed.1001953.
Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. ArXiv:1907.08228 [Cs]. http://arxiv.org/abs/1907.08228.
Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.
Skinner, B. F. (1963). Behaviorism at fifty. *Science*, *140*(3570), 951–958.
Skinner, B. F. (1976). *About behaviorism*. Vintage Books.
Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., … Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117.
Szot, T., Specht, C., Specht, M., & Dabrowski, P. S. (2019). Comparative analysis of positioning accuracy of Samsung Galaxy smartphones in stationary measurements. *PLoS ONE*, *14*(4), e0215562. https://doi.org/10.1371/journal.pone.0215562.
Taylor, P. J., Banks, F., Jolley, D., Ellis, D. A., Watson, S. J., Weiher, L., … Julku, J. (2021). Oral hygiene effects verbal and nonverbal displays of confidence. *Journal of Social Psychology*, *161*(2), 182–196. doi: 10.1080/00224545.2020.1784825.
Towse, J. N., Ellis, D. A., & Towse, A. S. (2020). Opening Pandora's Box: Peeking inside psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, *53*(4), 1455–1468. https://doi.org/10.3758/s13428-020-01486-1.
Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. MIT Press.
Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S. A., … Leow, A. (2018). Predicting mood disturbance severity with mobile phone keystroke metadata: A biaffect digital phenotyping study. *Journal of Medical Internet Research*, *20*(7), e241.

Generalizability, transferability, and the practice-to-practice gap

Joshua R. de Leeuw[a,*], Benjamin A. Motz[b,*], Emily R. Fyfe[b], Paulo F. Carvalho[c] and Robert L. Goldstone[b]

[a]Department of Cognitive Science, Vassar College, Poughkeepsie, NY 12604, USA; [b]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA and [c]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
jdeleeuw@vassar.edu; bmotz@indiana.edu; efyfe@indiana.edu; rgoldsto@indiana.edu; pcarvalh@andrew.cmu.edu; https://www.vassar.edu/faculty/jdeleeuw/; https://motzweb.sitehost.iu.edu/; https://psych.indiana.edu/directory/faculty/fyfe-emily.html; https://sites.google.com/view/paulocarvalho; https://psych.indiana.edu/directory/faculty/goldstone-robert.html

doi:10.1017/S0140525X21000406, e11

Abstract

Emphasizing the predictive success and practical utility of psychological science is an admirable goal but it will require a substantive shift in how we design research. Applied research often assumes that findings are transferable to all practices, insensitive to variation between implementations. We describe efforts to quantify and close this practice-to-practice gap in education research.

*Co-first authors.

Yarkoni's call for a focus on "predictive practical utility" led us to think about how scientists could adapt their methodological practices to meet this goal. One approach that a scientist might take is to shift toward applied work. For example, rather than running a learning experiment under the tight controls of a laboratory setting with the aim of establishing generalizable principles, researchers might instead run a learning experiment in a live classroom setting and test whether theoretical predictions improve educationally-relevant measures of student performance. Moving from the lab to the classroom (or any applied field) requires extensive revision to a study's structure, and will require researchers to specify, whether implicitly or explicitly, potentially relevant covariates that might otherwise be ignored. When translating from research-to-practice, these implementation variables could become useful signposts, informing where an intervention's benefits might extend. If this strategy were adopted collectively, fields might converge on reliable predictions about what interventions work in what contexts.

Unfortunately, we think that this description is more about our aspirations than our reality. While researchers now routinely run learning experiments in live classes (Motz, Carvalho, de Leeuw, & Goldstone, 2018), the predictions inferred from these studies are almost never informed by moderating variables (Koedinger, Booth, & Klahr, 2013). Studies conducted in small numbers of classes are commonly assumed to apply to all classes. Moreover, when field research observes null findings, the failure is often attributed to constraints of implementation rather than limitations of theory. As an example, classroom research on retrieval practice interventions is promising (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) but not consistently observed (e.g., Gurung & Burns, 2019; see also Moreira, Pinto, Starling, & Jaeger, 2019; Yang, Luo, Vadillo, Yu, & Shanks, 2021). Such mixed evidence hardly justifies the bold recommendations that it works for "all grade levels, all subject areas, and all students" (Agarwal, Roediger, McDaniel, & McDermott, 2020, p. 6). Sweeping practical recommendations like this are common in education; in some ways they constitute the very nature of the What Works Clearinghouse, a large evidence library of recommendations for how to intervene in education settings in useful ways, generalized from individual studies. Recommendations in education are no more reliable than the rest of psychological science, considering that two-thirds of US federally-funded impact studies found no impact (Schneider, 2018), and 50% of independent replication attempts in education fail to find evidence consistent with the original findings (Makel & Plucker, 2014). When researchers plan to intervene on the world, the crisis of generalizability is no less potent than it is for laboratory studies.

Given that applied research practices presently share most of the shortcomings of laboratory work with respect to generalizability, we predict that a shift toward testing theory in applied settings will not suffice. Even with such a focus, psychological scientists still exhibit a tendency to seek narrow, under-specified evidence of abstract principles, which are assumed to generalize across settings. By closing the research-to-practice gap, we do not necessarily close what we'll call the *practice-to-practice gap*: the benefits of an intervention, even when supported by field research that made accurate predictions for one practical setting, may not be transferable to other practical settings.

This concept of "transferability," more commonly associated with qualitative research, refers to the extent to which an intervention's effectiveness could be achieved in another sample and setting, whereas generalizability refers to the extent to which a sample statistic applies to the whole population and its many

situations. While synonymous with "replicability," transferability does not presume that an invariant global effect should exist in the first place (Lincoln & Guba, 1986), to be replicated or not. Rather, transferability presumes that an effect is conditionally dependent on context, analogous to state dependencies in a complex systems framework (Hawe, Shiell, & Riley, 2009).

Like Yarkoni, we believe that the transferability of an outcome in practice is contingent on variables which are typically not modeled, let alone articulated, in applied psychological science. Consider a teacher who hears that retrieval practice is an effective technique for improving student learning outcomes, and decides to incorporate regular practice quizzes into the curriculum. This teacher's implementation will likely deviate from the field tests where the technique was originally applied. As long as the field tests were carried out in narrow contexts, the teacher's deviations represent unmodeled sources of variance. Failing to account for this variance necessarily causes researchers to underestimate our uncertainty in the benefit of applying evidence-based practices to the teacher's classroom.

So what might an alternative approach look like? Yarkoni highlights the strategy of "design[ing] research with variation in mind," which raises the question: Which sources of variation? As Yarkoni points out, introducing variation makes the study significantly more resource-intensive to run. At some point adding additional sources of variation will have diminishing returns.

In the case of applied research, we think researchers could be guided by the natural variation that occurs in actual practice. This was the strategy for our *ManyClasses* study (Fyfe et al., 2021). We examined how the timing of feedback on student work affected learning performance in 38 different college classes. In each class, the experiment's parameters were allowed to vary according to instructors' preferences (difficulty, frequency, length, etc.). While the resulting set of implementations is by no means exhaustive, it does provide an estimate of the variance introduced when translating learning theory into normative instructional practice.

Purposefully introducing wide variation along theoretically important variables (e.g., Baribault et al., 2018) makes sense in laboratory contexts because this will assess generalizability to extreme and corner cases. In contrast, when concerned with the transferability of a phenomenon observed in the field, incorporating representative variation in settings will often be a better strategy if the goal is to determine how likely it is that the phenomenon will be observed in naturally occurring situations.

Financial support. Not applicable.

Conflict of interest. None.

References

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2020). *How to use retrieval practice to improve learning*. St. Louis, MO: Washington University in St. Louis. Retrieved from http://www.retrievalpractice.org.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Fyfe, E., de Leeuw, J. R., Carvalho, P. F., Goldstone, R., Sherman, J., Admiraal, D., … Motz, B. (2021). ManyClasses 1: Assessing the generalizable effect of immediate versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, *4*(3), 1–24. https://doi.org/10.1177/25152459211027575.

Gurung, R. A., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. https://doi.org/10.1002/acp.3507.

Hawe, P., Shiell, A., & Riley, T. (2009). Theorising interventions as events in systems. *American Journal of Community Psychology*, *43*(3–4), 267–276. https://doi.org/10.1007/s10464-009-9229-9.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937. https://doi.org/10.1126/science.1238056.

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? *Trustworthiness and Authenticity in Naturalistic evaluation. New Directions for Program Evaluation*, *1986*(30), 73–84.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513.

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*(5), 1–16. https://doi.org/10.3389/feduc.2019.00005.

Motz, B. A., Carvalho, P. F., de Leeuw, J. R., & Goldstone, R. L. (2018). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, *5*(2), 47–59. https://doi.org/10.18608/jla.2018.52.4.

Schneider, M. (2018). A more systematic approach to replicating research. IES Director's Blog, Institute of Education Sciences. Retrieved from https://ies.ed.gov/director/remarks/12-17-2018.asp.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309.

Lessons from behaviorism: The problem of construct-led science

Thomas E. Dickins[a] and Qazi Rahman[b]

[a]Department of Psychology, Middlesex University, London NW4 4BT, UK and [b]Institute of Psychiatry, Psychology & Neuroscience, Kings College London, Guy's Hospital, London SE1 9RT, UK.
t.dickins@mdx.ac.uk; http://tomdickins.net/
qazi.rahman@kcl.ac.uk; https://www.kcl.ac.uk/people/qazi-rahman

doi:10.1017/S0140525X2100008X, e12

Abstract

Yarkoni makes a number of valid points in his critical analysis of psychology, but he misses an opportunity to expose the root of its problems. That root is the poor practice around the derivation of explanatory constructs. We make comment on this with an example from behaviorist history and relate this to the recent discussion of scientific understanding in the philosophy of science.

For Yarkoni, the discipline of psychology suffers from at least two problems. First, the operationalized variables that are used in empirical work do not track the underlying structure of their hypotheses. This point has been made before with reference to failures to follow the hypothetico-deductive chain (Harris, 1976) and in a recent discussion of flexible versus hard to vary theories (Szollosi & Donkin, 2021). Second, the statistical assumptions made when using operationalized variables are in error.

We find ourselves in broad accord with Yarkoni's first diagnosis. But whilst he expresses agnosticism about psychological constructs, we believe construct formation is a cause of Yarkoni's problems. There is a deep history to be written about the use of constructs, but the case of behaviorism will help to make a point. Watson's original view was that only observable data could be included within an explanation of stimulus–response

transitions. However, when mathematical accounts proved untenable, this led to the introduction of unobservable constructs that were derived from observable data in order to generate an account. This was referred to as mediational neo-behaviorism (Moore, 2013). This derivation was from data collected in the laboratory; neo-behaviorism did not lead with the construct; it was not something to operationalize. Skinner noted that traditional psychology had a contrary practice, defining terms such as *memory* using unobservable constructs that were not derived from observable data (Skinner, 1945). He advocated looking to the reinforcement history of those terms within the discipline to understand what work they might be doing for scientists.

Skinner's point is related to Popper's discussion of definitions in science, in which he argued that the practice was to read definitions from left to right as <an x consists of p_1 to p_n properties> (Popper, 1945). This Aristotelian tradition introduces a form of essentialism, such that the project of science is to look for the essence of x. Instead, Popper claimed that definitions should instead be a form of shorthand. Once we understand that p_1 to p_n cohere in some way, for example, we can decide to name that kind of coherence x. Both Skinner and Popper committed to a clear-sighted form of empiricism.

It is often forgotten that behaviorism emerged as an antidote to introspection, which permitted verbal speculation about the architecture of internal behavioral causes. It was not that behaviorists denied inner experience, but they understood the scientific perils of trying to operationalize such models. Construct-led psychology necessarily has an introspective quality, and that practice leads to untethered ideas and a somewhat desperate attempt to empirically ground them. For the reasons that Skinner and Popper noted this will fail us scientifically: ideas, and more formally constructs, are best grounded when they emerge from empirical soil. Skinner also noted that the practices of cognitive psychologists were similar to behaviorists, in that they manipulated input variables and measured outputs, and were methodological behaviorists at best. Why not simply note regularities, titrate them, and then develop constructs? These points relate to Yarkoni's endorsement of a form of natural history.

De Regt claims that the unobservable mediational constructs that arose in behaviorism provided theoretical intelligibility, permitting the development of a functional explanatory framework that yielded prediction (de Regt, 2017). De Regt makes this more formal with his Criterion for Understanding Phenomena, which states that a phenomenon is understood if and only if it has an adequate explanation based on an intelligible theory. Furthermore, that theory must "conform to the basic epistemic values of empirical adequacy and internal consistency" (p. 92). Criteria for judging intelligibility include the ability of scientists to derive qualitative judgments about that theory without having to pursue exacting calculations. This package provides the necessary and sufficient conditions for scientific understanding. What we should note in the context of Yarkoni's argument is that here theory is being built in concert with empirical derivation, piece by piece.

Popper also revealed that there is no such thing as theory-free observation. Deciding what to measure is a theoretical choice and Yarkoni is well aware of this, and yet he avoids the discussion of grounding psychology in deeper theory. By this, we explicitly mean seeking some unity with biology, through the adoption of highly corroborated theories such as evolutionary theory, in order to provide a justifiable constraint on construct development. It is justified by the simple fact that behavioral plasticity is a phenotypic trait in the evolutionary framework (Meyers & Bull, 2002). In the last 30 years, this has been attempted by evolutionary psychology, but that particular exercise has sought to bolt traditional psychological constructs to hypotheses about adaptation, and to use adaptationist considerations to remove hypotheses that are not working (Dickins, 2003). A better use of evolutionary theory would be to adopt the optimality-led practices of behavioral ecology (Parker, 2006) and then look to develop constructs to explain internal causation of behavior, with a clear view of what behavior is for (cf. Curry, Mullins, & Whitehouse, 2019). Recent work taking a strong phylogenetic perspective on cognition, and borrowing from ecological psychology, which also had a distrust of construct-led science (Gibson, 1979), is carefully rebuilding the conceptual architecture of cognitive science (Bechtel & Bich, 2021). This work is cautious and thoroughly aware of all the assumptions it is making, building toward intelligible theory and understanding.

Yarkoni's statistical points about random and fixed effects are sound and we take his point that a portion of empirical psychology is really qualitative by nature. But we do not see the need to embrace this. Instead, in keeping with our recommendations above, we would advocate a stronger emphasis upon grounding psychology and deriving hypotheses from a biological "bottom up" – at least until workable idealizations of causation can be derived to allow future prediction (Potochnik, 2020). Doing this would introduce more steps into the derivation of hypotheses. This would include using modeling solutions to test the coherence of hypotheses.

Financial support. The preparation of this article was not supported by any research funding, but grew out of discussions between the two authors.

Conflict of interest. The authors declare no conflicts of interest.

References

Bechtel, W., & Bich, L. (2021). Grounding cognition: Heterarchical control mechanisms in biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376* (1820), 20190751. https://doi.org/10.1098/rstb.2019.0751.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478.

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

Dickins, T. E. (2003). What can evolutionary psychology tell us about cognitive architecture? *History and Philosophy of Psychology*, *5*(1), 1–16. http://dspace.uel.ac.uk/jspui/handle/10552/566.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin Company.

Harris, R. J. (1976). The uncertain connection between verbal theories and research hypotheses in social psychology. *Journal of Experimental Social Psychology*, *12*(2), 210–219. https://doi.org/10.1016/0022-1031(76)90071-8.

Meyers, L. A., & Bull, J. J. (2002). Fighting change with change: Adaptive variation in an uncertain world. *TRENDS in Ecology and Evolution*, *17*(12), 551–557. http://dialnet.unirioja.es/servlet/articulo?codigo=2851914%5Cnpapers://994295af-42ed-47fe-9a2a-07e2e7d6a6b8/Paper/p709.

Moore, J. (2013). Three views of behaviorism. *Psychological Record*, *63*(3), 681–692. https://doi.org/10.11133/j.tpr.2013.63.3.020.

Parker, G. A. (2006). Behavioural ecology: Natural history as science. In J. Lucas & L. Simmons (Eds.), *Essays in animal behaviour* (pp. 23–56). Academic Press.

Popper, K. R. (1945). *The open society and its enemies*. Routledge.

Potochnik, A. (2020). Idealization and many aims. *Philosophy of Science*, *87*(5), 933–943. https://doi.org/10.1086/710622.

Skinner, B. F. (1945). Analysis of psychological terms. *Psychological Review*, *52*, 270–277.

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796.

Observing effects in various contexts won't give us general psychological theories

Chris Donkin[a], Aba Szollosi[b] and Neil R. Bramley[b]

[a]School of Psychology, University of New South Wales, 2052 Sydney, Australia and [b]School of Philosophy, Psychology and Language Sciences, Edinburgh EH8, Scotland

christopher.donkin@gmail.com, aba.szollosi@gmail.com, neil.bramley@ed.ac.uk

doi:10.1017/S0140525X21000479, e13

Abstract

Generalization does not come from repeatedly observing phenomena in numerous settings, but from theories explaining what is general in those phenomena. Expecting future behavior to look like past observations is especially problematic in psychology, where behaviors change when people's knowledge changes. Psychology should thus focus on theories of people's capacity to create and apply new representations of their environments.

Generalization is inherent to scientific inference – physicists learn about the properties of distant stars based on lab-based experiments, vision scientists use illusions to infer mechanisms of basic human perception, and meteorologists generalize from the simulations of mathematical models to form expectations about tomorrow's weather. How do scientists draw conclusions about the general features of the environment using such incomplete and indirect evidence?

The position taken in the target article is that generalization is licensed by what we have observed. In this view, we learn about the robustness of a phenomenon by observing it in a range of contexts, and on that basis we can generalize it to a wider range of environments (i.e., expect the phenomenon to occur in those other environments similarly). However, as Yarkoni demonstrates in his Section 4, the choice of any specific generalization that goes beyond what has been directly observed is arbitrary and irrational (see also Hume, 1739). Unfortunately, it seems that Yarkoni does not take his own critique seriously, and so misses the opportunity to explain why generalization works.

What is absent from the above description is the role theory plays in generalization. Scientific theories make claims about how our physical environment behaves and why, and so they imply what features of these environments generalize (Deutsch, 2011). That is, rather than on the basis of repeated observations, we expect to see a phenomenon generalize to a new context because a theory implies that we should.

Take the Stroop effect, widely considered to be one of the most general findings in psychological science. The robustness of the Stroop effect comes not from having observed it repeatedly and in many contexts, but because we have a compelling explanation for why it happens. Namely, the processes of reading and naming both rely on the same semantic representation. Also, quickly reading the words we see is a well-established habit, while naming the color in which a word is printed is unusual and thus slower. Therefore, if reading a word creates semantic activation that could be mistaken as the output of naming, the word is spoken preemptively. Such an explanation dictates the kinds of experiments in which we should observe an effect – for example, if we encourage reading by making most words either helpful (the same color as the ink) or benign (words unlikely to activate concepts related to colors), and it helps us avoid experiments likely to fail – such as printing the color word in a language foreign to the reader. Thus it is the general implications of this explanation that allow us to produce the Stroop effect in so many contexts.

The generality implied by theories, however, should not be taken for granted. We should not, for example, take seriously any specific prediction about generality coming from flexible theories, since they could be easily changed to predict any possible result (Szollosi & Donkin, 2021). Similarly, most phenomena in psychology are explained by a number of possible theories, with no clear reasons to decide between them. Where such theories diverge in their claims about generality, it would be entirely arbitrary to expect the predictions of any one theory. So, while Yarkoni is right to be concerned about generalizations in psychology research, the real problem is that generalization is expected to occur despite having no good reason to expect it over the countless other (and often non-general) explanations that are at least equally (and often more) viable.

Generalization in psychology is also complicated by the fact that what is often considered its explicanda – people's thoughts, motives, and behaviors – all tend to change in response to new knowledge. For example, the effectiveness of a demonstration of the Stroop effect will likely diminish as the task is repeated. Once aware of the effect, a participant may try to explain what is happening, and come up with strategies to defeat it, such as only looking at a small part of the text or unfocusing their eyes.

More generally, when placed into an unfamiliar setting or environment, people are liable to adapt, combine, and repurpose (aspects of) their existing knowledge until they find an adequate representation or explanation (e.g., Bramley, Dayan, Griffiths, & Lagnado, 2017; Lake, Ullman, Tenenbaum, & Gershman, 2017; Szollosi & Newell, 2020). This new or restructured knowledge may permit behaviors that are entirely novel, and could never have been predicted (else that knowledge already existed). Thus, since the most characteristically human aspects of behavior are driven by representations and these are always subject to change, it is unreasonable to expect any single observed behavior to be completely general. Instead, psychology should seek the causes of this flexibility and adaptivity in observed behaviors (Chomsky, 1959).

Therefore, to understand what can be generally true of cognition, we should explain how people can create and change their representations of their surroundings. This suggests that primary explicanda of psychology are people's *capacities*, not any particular behavior (e.g., van Rooij & Baggio, 2021). Psychological explanations should not only account for what people did in some experiment, but also for what they could have done. In explaining the Stroop effect, a capacity-focused explanation must explain not only why people exhibit the Stroop effect when first exposed, but also reasons why they might not, or why the effect might change in the future (e.g., if the Stroop effect were to be taught in high schools, to the point that people learn strategies to avoid it).

Understanding the general aspects of human behavior by collecting a list of psychological effects and the contexts in which

they occur – as suggested in the target article – is incompatible with the fact that we should expect behavior *not* to be the same across contexts. It is not clear how the effect-focused psychological science for which Yarkoni advocates could ever be reconciled with the flexibility of human cognition. Instead, we suggest that a focus on explaining people's capacity – in terms of a general ability to create and adapt representations of environments – could result in stronger, broader psychological generalizations.

Financial support. Chris Donkin was funded by an ARC Discovery grant (DP190101675). Aba Szollosi and Neil R. Bramley were funded by an EPSRC New Investigator grant (EP/T033967/1).

Conflict of interest. None.

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. https://doi.org/10.1037/rev0000061

Chomsky, N. (1959). Review of B. F. Skinner's Verbal Behaviour. *Language*, *35*(1), 26–58. https://doi.org/10.2307/411334

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. Allen Lane.

Hume, D. (1739). *A treatise of human nature*. London: John Noon.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between confirmatory and exploratory research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008–1018. https://doi.org/10.1016/j.tics.2020.09.005

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Addressing a crisis of generalizability with large-scale construct validation

Jessica Kay Flake, Raymond Luong and Mairead Shaw

Department of Psychology, McGill University, Montreal, QC H3A 1G1, Canada.
Jessica.flake@mcgill.ca; raymond.luong@mail.mcgill.ca;
mairead.shaw@mail.mcgill.ca;
https://www.mcgill.ca/psychology/jessica-kay-flake;
https://www.mcgill.ca/psychology/jessica-kay-flake

doi:10.1017/S0140525X21000376, e14

Abstract

Because of the misspecification of models and specificity of operationalizations, many studies produce claims of limited utility. We suggest a path forward that requires taking a few steps back. Researchers can retool large-scale replications to conduct the descriptive research which assesses the generalizability of constructs. Large-scale construct validation is feasible and a necessary next step in addressing the generalizability crisis.

What Yarkoni describes is grim: any statistical model estimated from any study has so many omitted sources of variance that the estimates are likely meaningless. The gross misspecification of our models and specificity of our operationalizations produce claims with generality so narrow that no one would be interested in them. From this, one could reasonably conclude that a single laboratory of researchers should struggle to design an experiment worth the time it takes to carry it out.

For those who agree – and we count ourselves among them – what are the possible steps forward? There are seemingly infinite sources of invalidity in our work; where do we start? Of the solutions Yarkoni describes, we expand on ideas of large-scale descriptive research that are feasible and worthwhile.

The foundational assumption of construct validity and generalizability

Construct validity is a linchpin in the research process. When researchers create numbers from measurements, it is assumed those numbers take on the intended meaning. A foundational challenge for psychological scientists is ensuring this assumption holds so that those numbers are valid and that their meaning generalizes across the range of interpretations made about them. When psychologists study constructs like motivation, personality, and individualism, they don't intend to only describe the people in their sample, they intend to describe something meaningful and global about the human condition.

Psychometricians refer to the evaluation of the assumption of construct validity and construct generalizability as on-going construct validation (Cronbach & Meehl, 1955; Kane, 2013). On-going construct validation is possible with classic and modern psychometric methods for many approaches to measurement that are common in psychology. For example, the Trends in International Mathematics and Science Study measures mathematics and science achievement of children from 60 countries and is the culmination of years of quantitative and qualitative research to determine how to measure such constructs and generate valid scores that are comparable across diverse peoples (TIMSS; https://www.iea.nl/studies/iea/timss). A concrete step forward is for psychologists in other areas of study to take what they are measuring and the generalizability of what they are measuring as seriously as the scientists who created TIMSS take achievement.

However, this step comes well before conducting studies using these measures to test relationships and causal effects. It requires work that psychology has historically undervalued: systematic review and synthesis of theory with an emphasis on organizing old ideas instead of generating new ones, mixed methods, representative sampling, and descriptive research on constructs and the variability in scoring that measuring them in different contexts can cause.

What does large-scale construct validation research look like?

Psychology knows what large-scale collaborative studies look like because large-scale replications have become a norm following the Reproducibility Project: Psychology (Open Science Collaboration, 2015). The Many Labs collaboration is on its fifth iteration (ML; Ebersole et al., 2020) and published registered replication reports typically include data collection efforts spanning across dozens of laboratories (e.g., Wagenmakers et al., 2016). However, these replication studies skip over construct validation. We reviewed the

measures used in ML2 and found that surveys collected to measure key variables of interest in the replication study had limited validity evidence and poor psychometric properties. For example, a measure of subjective well-being was used in a replication of an effect reported by Anderson, Kraus, Galinsky, and Keltner (2012). We tested the assumed factor structure of this scale using ML2 data and by any conventional standards the model fit was poor (CFI = 0.616, RMSEA = 0.262, and SRMR = 0.267; Shaw, Cloos, Luong, Elbaz, & Flake, 2020), casting doubt that the scores represented well-being. How can we interpret the replicability of an effect if the numbers used in the analysis don't have the meaning the original researchers intended? The results of this review are consistent with what Yarkoni is saying: psychologists have bent over backwards trying to replicate effects that didn't convey anything meaningful in the first place.

Luckily, we can squeeze more juice from completed large-scale replication studies with post hoc construct validation. For example, Cloos and Flake (submitted) assessed the psychometric properties of an instrument used in ML2 and if those properties generalized across two translated versions. The short story is that a critical unmodeled source of variance in replication results is measurement heterogeneity introduced by translation. Researchers could also take this approach with single studies that are published with materials and data. If the instruments from single studies could also be systematically reviewed, reanalyzed, and synthesized, psychologists could generate compelling evidence for the generalizability of constructs. This is something researchers can retroactively work on moving forward. But this is not an efficient process. Ideally studies would step back from replicating effects and focus on the theoretical merit and measurement approaches of key constructs in a fashion that exposes them to substantial heterogeneity (e.g., across data collection settings, time, and cultures). The constructs and associated measures that demonstrate validity and generalizability are then good candidates for replication studies.

We are currently attempting a version of this with the Psychological Science Accelerator (Moshontz et al., 2018). We are taking two measures originally developed in English and evaluating validity and generalizability in over 20 languages. We aren't testing any key effects; we are just going to describe the measures and their properties across a diverse set of languages using a mixed-methods approach. Regardless of the results, we will generate useful knowledge about these constructs and the feasibility of using existing measures to study them on a global scale.

Large-scale construct validation is methodologically and logistically challenging with very few incentives for doing it. An optimistic interpretation is that there is plenty to do. If we focus less on generating new ideas and more on organizing, synthesizing, measuring, and assessing constructs from existing ideas, we could keep busy for decades. Probably longer, in fact! Scientists spent literally hundreds of years determining what an electric charge was, the units it should be measured in, and how to measure it. As researchers in a far younger discipline with abstract and unobservable constructs, those hundreds of years are likely still ahead of us.

Acknowledgments. None.

Financial support. Jessica Kay Flake and Raymond Luong's work was funded by the Ware's Prospector's Innovation Fund awarded to Jessica Kay Flake (249040, 2019). Mairead Shaw's work was funded by Fonds de recherche du Quebec – Nature et technologies awarded to Mairead Shaw (288759, 2020)

Conflict of interest. On multiple occasions we have gotten ice cream with the author of the target article.

References

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect: Social status and subjective well-being. *Psychological Science*, *23*(7), 764–771. https://doi.org/10.1177/0956797611434537.

Cloos, L. J. R., & Flake, J. K. (submitted). *Lost in translation? Addressing measurement equivalence in large-scale replications*. Department of Psychology, McGill University.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957.

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., … Nosek, B. A. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *The Journal of Educational Measurement*, *50*(1), 1–73.

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716.

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications : Insights from Many Labs 2. *Canadian Psychology*, *61*(4), 289–298. https://doi.org/10.1037/cap0000220.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458.

Mismatch between scientific theories and statistical models

Andrew Gelman

Department of Statistics, Columbia University, New York, NY 10027, USA.
gelman@stat.columbia.edu; http://www.stat.columbia.edu/~gelman/

doi:10.1017/S0140525X21000091, e15

Abstract

Yarkoni recommends that psychology researchers should take care to align their statistical models to the verbal theories they are studying and testing. This principle applies not just to qualitative theories in psychology but also to more quantitative sciences: there, too, mismatch between open-ended theories and specific statistical models have led to confusion.

In this comment, I would like to first put Yarkoni's paper in the context of statistical reasoning and then illustrate that the problem he discusses arises not just in psychology but in other sciences as well.

Following Popper (1934/1959) and Lakatos (1978), we can consider two basic paradigms of scientific inference:

(1) *Confirmation:* You gather data and look for evidence in support of your research hypothesis. This could be done in various ways, but one standard approach is via statistical significance testing: the goal is to reject a null hypothesis, and then this rejection will supply evidence in favor of your preferred research hypothesis.
(2) *Falsification:* You use your research hypothesis to make specific (probabilistic) predictions and then gather data and perform analyses with the goal of rejecting your hypothesis.

It is tempting to consider confirmationist reasoning as bad and falsificationist reasoning as good, but both have their role within good scientific practice. Mayo (1996) considers these inferential approaches as part of a larger process in which experiments are designed to test and adjudicate among competing hypotheses.

It is important to distinguish these from a third, erroneous mode of reasoning:

(3) *Naive confirmationism*: You start with scientific hypothesis A, then as a way of confirming this hypothesis, the researcher comes up with null hypothesis B. Data are found which reject B, and this is taken as evidence in support of A.

In Yarkoni's terminology, hypothesis A is a verbal assertion in psychology, and null hypothesis B is a statistical model. When expressed above, naive confirmationism is an obvious logical fallacy, but it is done all the time in research published in top journals. For example, Durante, Arsena, and Griskevicius (2013) used survey data to claim that "the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships" offering as evidential support the rejection of a series of statistical null hypotheses.

The difficulty here is that either the scientific hypothesis is general and non-quantitative (in which case, sure, the ovulatory cycle, like everything else, will have *some* nonzero effect, and so the confirmation of this vapid hypothesis tells us nothing whatsoever) or the hypothesis is quantitative – what are the purported effects, how large are they, and how persistent are they across people and across settings – in which case these effects need to be studied with a statistical model appropriate to the task, and the rejection of an empty null is irrelevant. See Gelman (2015) for further discussion of this example in the general context of varying treatment effects.

In his article, Yarkoni focuses on verbal hypotheses in psychology, but similar problems arise in other fields. For example, Chen et al. (2013) claim that a coal-heating policy caused a reduction of life expectancy of 5.1 years in half of China, with the supporting evidence coming from a discontinuity regression. In this case, the scientific model is quantitative, but there is still a disconnect with the statistical model, so that the empirical claims are questionable: to put it in statistical terms, the 95% confidence intervals do not have 95% coverage, and the null hypothesis can reject much more than 5% of the time even if there is no effect (Gelman & Imbens, 2019; Gelman & Zelizer, 2015). The problem is that the statistical model makes many assumptions beyond the scientific model of the effect of pollution.

It would be usual to characterize the two above stories as statistical errors. In the ovulation example, the mistake is to make a strong conclusion from the rejection of a null hypothesis, without recognizing that in practice all statistical hypotheses are false; and in the coal-heating example, the mistake is to use a flawed statistical model that overfits to patterns in the data which are not pure noise (for that, the regression would indeed have its advertised statistical properties). In addition, both cases are examples of the garden of forking paths, by which an analyst can choose among many possible statistical tests to apply to a problem, making it easier to obtain statistical significance and thus publishable results.

Following Yarkoni, though, we see these not just as examples of poor analysis or questionable research practices, but as special cases of the general problem of mismatch between scientific and statistical models.

Moving forward, we should recognize the limitations of any statistical model – and I say this as a person who constructs and fits these models for a living. Rather than thinking of "the model" or "the test" corresponding to a scientific or engineering question, we recommend fitting a multitude of models. Think of models as tools. When building a house, or even simply installing a shelf, we don't just use a hammer or a screwdriver or a level. We use all these tools and more. Different models and statistical tests capture different aspects of the data we observe and the underlying structure we are trying to study.

Financial support. U.S. Office of Naval Research.

Conflict of interest. None.

References

Chen, Y., Ebenstein, A., Greenstone, M., & Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences*, *110*, 12936–12941.

Durante, K. M., Arsena, A. R., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*, 1007–1016.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*, 632–643.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, *37*, 447–456.

Gelman, A., & Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research and Politics*, *2*, 1–7.

Lakatos, I. (1978). *Philosophical papers*. Cambridge University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Popper, K. R. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.

We need to think more about how we conduct research

Gerd Gigerenzer

Max Planck Institute for Human Development, Berlin Lentzeallee 94, 14195 Berlin, Germany.
gigerenzer@mpib-berlin.mpg.de

doi:10.1017/S0140525X21000327, e16

Abstract

Research practice is too often shaped by routines rather than reflection. The routine of sampling subjects, but not stimuli, is a case in point, leading to unwarranted generalizations. It likely originated out of administrative rather than scientific concerns. The routine of sampling subjects and testing their averages for significance is reinforced by delusions about its meaningfulness, including the replicability delusion.

The replicability crisis has made us rethink research practice. Should we lower the level of significance from 0.05 to 0.005, replace *p*-values with Bayes-factors, or require preregistration? Yarkoni rightly advocates looking even deeper into what fuels unwarranted generalizations.

Select or sample stimuli?

As Yarkoni notes, the routine practice is to sample subjects but not stimuli, although the choice of stimuli can more strongly influence a result. This happens if individuals are more alike than stimuli or if the stimuli are not representative – akin to a survey reporting only the extreme opinions of a few selected people rather than those of a representative sample (Brunswik, 1956). Consider two prominent cases where generalizations based on selected stimuli become invalid.

Take the claim that people are overconfident. It is based, among others, on a large number of studies asking general knowledge questions such as "Which city lies further south: Rome or New York? How confident are you?" On average, confidence was higher than proportion correct. Rome, however, is further north than New York yet warmer. Selecting unusual stimuli generates a semblance of general overconfidence. When stimuli were instead randomly sampled from a population (such as all large cities in the world), average confidence matched proportion correct (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olssen, 2000).

The same holds for the generalization that availability makes people overestimate the likelihood that a letter more likely occurs first than third in a word. In the original study (Tversky & Kahneman, 1973), two-thirds overestimated this for the five consonants selected – all of which more likely appear in the third position, which is untypical. But when representative samples of all letters were used, judgments corresponded to the actual letter frequencies rather than to availability (Sedlmeier, Hertwig, & Gigerenzer, 1998).

In both cases, the sampling of stimuli, not subjects, made the essential difference. By picking unusual stimuli, one can produce results that do not generalize.

Why sample subjects but not stimuli?

One might object that sampling subjects is, and has always been, the method of experimental psychology since its beginnings in Wundt's laboratory. In fact, research practice consisted of careful analysis of single individuals exposed to many stimuli. Wundt himself served as a subject, tested by a technician on a range of stimuli. Luria studied the mind of the mnemonist Shereshevsky using a broad range of stimuli, including words, numbers, and tones. Skinner studied one pigeon at a time, reporting cumulative records instead of averages. Simon studied individual chess players, varying chess positions.

One might also object that sampling subjects is required by inferential statistics. That is also incorrect. Take Fisher's *Design of Experiments* (1935), which introduced psychologists to randomized experiments, null hypotheses, and significance. It reported one psychological experiment involving a single subject (a lady who allegedly knew whether tea or milk was first poured into a cup) and a sample of stimuli (cups of tea). Nowhere in Fisher's experiments were subjects ever sampled (Gigerenzer, 2006). Why did research practice change?

In his seminal book *Constructing the Subject*, Danziger (1990) argues that American psychologists' reason for abandoning carefully study of individuals in the 1930s and 1940s and embracing averages as their new "subject" had little to do with science. They reacted to university administrators' pressure to prove that their research was useful for applied fields, specifically educational research, which offered large sources of funding. The use of averages in treatment groups (such as pupils) was quickly adopted in educational psychology and parapsychology, whilst the core of scientific psychology, such as research on perception research, continued to conduct studies with few individuals. If Danziger is right, our current research practice – sampling subjects only, and testing their averages for

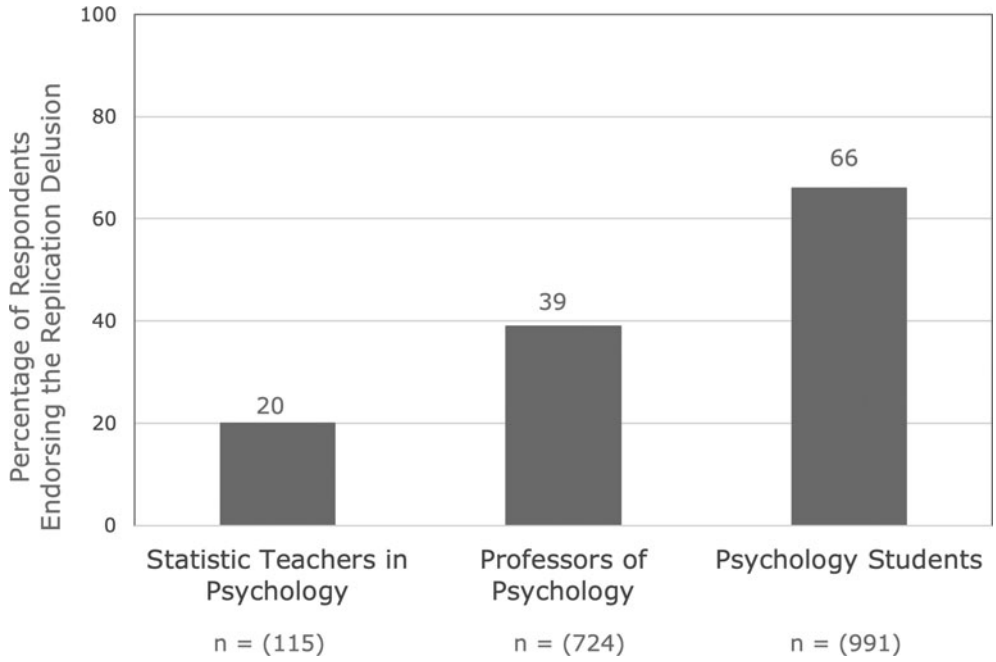

Figure 1. (Gigerenzer) The replication delusion. Shown is the percentage of respondents who endorsed that $p = 0.01$ logically implies a 99% chance of replication. For details, see Gigerenzer (2018).

significance – is a historical artifact motivated by the needs of administrators, not science.

The replication delusion

Routines easily turn into blind spots, which also exist in the sampling of subjects. The statistical theory underlying significance testing assumes that random samples are drawn from a population, yet most researchers do not sample subjects (or stimuli) randomly from a population or define a population in the first place. Without meeting the assumptions of the model, one cannot know the population to which a significant result might generalize, or where it might be replicated. That fundamental mismatch between statistical theory and experimental practice is bridged by a delusion:

> A significant result $p = 0.01$ logically implies that if the experiment were repeated a large number of times, one would expect to obtain a significant result on 99% of occasions.

This delusion provides unwarranted certainty that a result is replicable, and makes replication studies appear obsolete – and not worth publishing. How widespread is it? In all existing studies, mostly conducted in the last decade, 839 academic psychologists in Chile, Germany, Italy, the Netherlands, Spain, and the UK had been asked to evaluate the above statement – the replication delusion – including variants such as $p = 0.001$, which made no difference (Gigerenzer, 2018) (Fig. 1).

All in all, a total of 20% of the faculty teaching statistics in psychology departments considered the statement correct, as did 39% of professors of psychology and lecturers, and, unsurprisingly, two-thirds of 991 students.

The replication delusion is not alone in maintaining the belief in null hypothesis testing as a universal method. There is also the *illusion of certainty* (that significance disproves the null hypothesis and non-significance proves the experimental hypothesis) and *Bayesian wishful thinking* (that the *p*-value determines the probability of the null hypothesis or of the experimental hypothesis). In every study, the majority of researchers (56–97%) exhibited one or more delusions about what a significant *p*-value means (Gigerenzer, 2018).

All this leads to a positive conclusion: We should use the momentum of the replicability crisis to liberate research practice from methodological rituals and associated delusions, and we might conduct research to find out why we do what we do.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge, UK: Cambridge University Press.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Gigerenzer, G. (2006). What's in a sample? A manual for building cognitive theories. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 239–260). New York: Cambridge University Press.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*, 198–218. doi: 10.1177/2515245918771329.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528. doi: 10.1037/0033-295X.98.4.506.

Juslin, P., Winman, A., & Olssen, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396. doi: 10.1037/0033-295X.107.2.384.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 754–770. doi: 10.1037/0278-7393.24.3.754.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

The four different modes of psychological explanation, and their proper evaluative schemas

Michael Gilead

Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel.
michael.gilead@gmail.com

doi:10.1017/S0140525X21000145, e17

Abstract

I apply Benjamin's (1941) taxonomy of common scientific "modes of explanation" to the psychological context. I argue that: (i) in a "naming" mode, generalizability is not necessary; (ii) in an "analysis," generalizability is desired; (iii) in a "causal ontology," generalizability is merely one of the means to an end (theory-challenge); (iv) in a "synthesis," generalizability is (eventually) critical. A better appreciation of the diversity in psychologists' modes of explanation is crucial for cogent meta-psychological discussions.

I argue that the underlying problem reflected in the target article is that psychologists often apply inappropriate schemas when evaluating our own research. Applying the wrong evaluative schema can cause us to ignore questions of generalizability when they are crucial – but can also cause us to emphasize generalizability when it is of secondary importance.

In order to explicate my reasoning, I reinterpret Benjamin's (1941) taxonomy of common scientific "modes of explanation." As I argue below, the issue of generalizability should be treated differently in the different modes.

Phase I: name

Sometimes, scientists simply identify and name a phenomenon, and provide us with a prescription for using their neologism ourselves. As an example, consider the effect of "hedonic adaptation" – the observation that individuals may adapt to an improvement\decrement in their life circumstances, and return to baseline levels of well-being (Frederick & Loewenstein, 1999).

In this "naming" mode (prevalent in social psychology) it is not crucial to demonstrate that the phenomenon is omnipresent. It is informative if some of the people some of the time exhibit "hedonic adaptation," "ingroup bias," or "pluralistic ignorance." Surely, we will be interested to know who exhibits the phenomenon and when, but this is a matter for the next phase of the scientific process.

Phase II: analysis\description

Once we have identified an entity, we may start to characterize its properties. For example, the relation between A (e.g., Conservatism) and B (Happiness) has form C and is moderated by D. A scientist can advance our understanding by finding such regularities – even without providing us with a full-fledged theory.

To provide such analyses, scientists rely on induction. Principled induction entails specifying a model embodying our assumptions regarding the possible nature of variability in the world and our study; when we omit sources of variability, we sweep them under the *ceteris paribus* rug (e.g., assuming stimuli will work the same, people will act the same). Sources of variability are infinite, and cannot be fully estimated or cogently assumed away. This means that all inductions entail a leap of faith – and are considered logically invalid (e.g., Hume, 2003).

However, not all inductions are equal. As our observations become more comprehensive, and when our leaps of faith are relatively cogent, our inductions become better – in the sense that they are more likely to hold in novel contexts. Yarkoni is clearly right in saying that we'd be happy to find very general laws, or at least laws that apply to some well-known scope.

Nonetheless, I think Yarkoni is mistaken if he is saying that way to evaluate *any* research that generalizes is by assessing the congruence between the scientists' summary of their results, and what the results "actually" show (i.e., assessing the strength of their "inductive argument"). Litigating the congruency of evidence against belief is a proper way to test bona fide theories (see below). It is the wrong schema for evaluation of purely descriptive generalizations, because the generalizability of a pattern is fully independent of the claims made about its generalizability. In a research paper, we (should) have what we need to independently assess the appropriate scope of generalization (indeed, the target article makes such assessments); therefore, it is irrelevant whether an author is grandstanding.

Surely, as reviewers, we should tell scientists to "tone it down" when they overstate their results. However, science shouldn't be personal – we should not confuse rhetoric/aesthetics/ethics with epistemology.

Phase III: causal ontology

Once we observed the properties of entities, we can generate claims about causes that may have given rise to these properties. For example, we may argue that people's evaluation of an event is most affected by its ending (i.e., the "end rule," Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993) because people better remember recent events (Baddeley & Hitch, 1993).

Causes are never "out there" in the world and thus must be imagined rather than directly observed (Kant, 1908). After we have imagined a causal theory, we try to empirically justify it – and can often do so by deducing and testing the theory's entailments. One of the ways through which scientists justify their theories is by fiercely challenging them, giving rise to gradual "survival of the fittest" (Popper, 1999).

Theory-challenge can be a lengthy process, pursued through various routes. We may seek the boundaries of a theory by testing it on broad, representative samples of individuals and stimuli. However, oftentimes, a good way to challenge a theory is by testing it in narrow, unusual contexts (e.g., does the "peak-end" effect occur in people with episodic amnesia?). Thus, gathering evidence for broad applicability is (one of the) means for severe theory-challenging – not its end-goal.

Phase IV: synthesis

Once we have some ontology of causes, we can seek a full-fledged "synthesis" that explains how a phenomenon can be truly accounted for in terms of a set of causes and their relations. For example, Rutledge, Skandali, Dayan, and Dolan (2014) showed how people's momentary well-being is captured by a formula that describes precise relations between causes such as "prediction error" and "expected rewards."

The typical way to conduct such research (prevalent in cognitivist "computational modeling" studies) is to pick an operationalization of a phenomenon, and gradually try to find the model that best fits the observed data. As such, those who conduct such research can pride themselves in severely challenging their models.

A caveat of such research is that the strong focus on a specific paradigm as a benchmark often leads us to forget that operationalization was once merely a means to study the general phenomenon. We should (but often forget to) ask the question raised by Yarkoni – can our model generalize to additional manifestations of the phenomenon? However, we must also remember that scientific progress can be slow; it is possible that models that explain behavior on a specific paradigm will eventually develop into full, generalizable accounts of a phenomenon.

Summary

The abovementioned modes of explanation can all reduce puzzlement. A better appreciation of these different modes is critical for cogent discussions concerning replicability, generalizability, and the utility of psychological science.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Baddeley, A. D., & Hitch, G. (1993). The recency effect – implicit learning with explicit retrieval. *Memory & Cognition*, *21*(2), 146–155. doi:10.3758/bf03202726.

Benjamin, A. C. (1941). Modes of scientific explanation. *Philosophy of Science*, *8*(4), 486–492.

Frederick, S., & Loewenstein, G. (1999). 16 Hedonic adaptation. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *The foundations of Hedonic* (pp. 302–329). Russell Sage.

Hume, D. (2003). *A treatise of human nature*. Courier Corporation.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405.

Kant, I. (1908). Critique of pure reason. 1781. *Modern classical philosophers* (pp. 370–456). Houghton Mifflin.

Popper, K. R. (1999). *All life is problem solving*. Psychology Press.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences of the USA*, *111*(33), 12252–12257. doi:10.1073/pnas.1407535111.

The cost of crisis in clinical psychological science

Joshua B. Grubbs

Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, USA.
GrubbsJ@BGSU.edu; www.JoshuaGrubbsPhD.com

doi:10.1017/S0140525X21000388, e18

Abstract

Yarkoni has argued that psychology is facing a generalizability crisis, but the real cost of this crisis is obscured by a focus on topics from psychology's most academic subfields. Psychology is also filled with applied subfields, and it is within those subfields – especially clinical science – where the cost of a generalizability crisis will be most severe.

The past decade has clearly demonstrated that psychology is replete with contradictory, redundant, and irreplicable findings. Ego depletion exists (Baumeister & Vohs, 2007), except when it doesn't (Hagger et al., 2016), or when it's too small to have a meaningful effect (Dang et al., 2021). Depending on which psychologist is writing the paper, people may alternatively have grit (Duckworth & Quinn, 2009), be resilient (Friborg, Barlaug, Martinussen, Rosenvinge, & Hjemdal, 2005), or just demonstrate specific aspects of conscientiousness (Schmidt, Nagy, Fleckenstein, Möller, & Retelsdorf, 2018). To an objective observer, it may seem that large-swaths of psychology have long moved past Feynman's (1974) warnings of a cargo cult science and all-but-abandoned the science part altogether. As more psychologists have publicly acknowledged that huge portions of our science are merely either the selling of old wine in new bottles or attempts to sell bottled water by calling it wine, the past decade in psychological science has been a crisis of crises. The replication crisis has given way to a measurement crisis (Flake & Fried, 2020), a validity crisis (Clark & Watson, 2019; Flake, Pek, & Hehman, 2017; Lundh, 2019), a practicality crisis (Berkman & Wilson, 2021), and a litany of others (Hughes, 2018). To this burgeoning but rich legacy of crises, Yarkoni has added another: The generalizability crisis.

Yarkoni has argued that psychology has fallen into what Meehl (1990) warned of, having created a field of scientists whose entire careers are devoted to describing, demonstrating, and marketing nothing. This, in turn, wastes resources such as time, effort, and funds, as all are being spent on experimental studies that have no chance of actually answering the question they purport to answer. Perhaps more concerningly, this crisis points to a loss of credibility for the entire field. Indeed, these costs are high. Yet, for some domains within psychology, a crisis of generalizability may come at a greater cost than wasted resources or statistical charades.

For seminal work in verbal overshadowing to fall short of basic generalizability is disappointing, particularly given the efforts invested in multisite replication efforts. Even so, would anyone argue that social costs of these efforts extend much further beyond the loss of those resources? Similarly, the three-decade-long-and-still-continuing obsession with ego depletion has undoubtedly done more to deplete financial and human capital resources than demonstrate any generalizable claim about willpower or self-control (Friese, Loschelder, Gieseler, Frankenbach, & Inzlicht, 2019). Yet the societal costs of those Sisyphean attempts are likely relatively low. Failed experiments certainly may have derailed careers, and such losses should not be ignored. Even so, the average layperson will be more influenced by the ice-cream they choose to eat for dessert today than they will by failure of ego depletion to deliver on its years of bold claims. Could the same be said for interventions in clinical psychology? If evidence-based treatments and "gold-standard" therapies for depression, post-traumatic stress, or anorexia lack generalizability, are the costs accurately measured in wasted scientific resources?

Even though all applied scientists in psychology should be deeply concerned about the cost of a generalizability crisis, clinical science should be so in even greater measure. It is no accident that Meehl's perennially salient concerns about the quality and rigor of psychological science were born of a career that integrated clinical practice (Meehl, 1987). The costs of ungeneralizable and sloppy science are higher when that science is guiding practitioners' decisions. There is abundant evidence that many of clinical psychology's gold-standard treatments are built on statistical errors, questionable research practices, and bad inferences (Sakaluk, Williams, Kilshaw, & Rhyner, 2019). If even a few of the remaining treatments that are not built on gross statistical errors or questionable research practices are found to lack generalizability – a critique that practicing clinicians have voiced for many years (Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2013; Stewart, Stirman, & Chambless, 2012) – what then remains of the field?

Undoubtedly, the same criticisms and concerns that I have framed around clinical psychology might also apply to other applied domains. Educational psychology has influenced real-world decisions for years, and the work done in industrial and organizational psychology has affected careers of countless individuals. Similarly, there are social and cognitive psychologists that are heavily engaged in applied research that bears real-world implications for both individual people and broader policy. Such fields and scientists should be similarly shaken by the notion that their work fundamentally does not and cannot actually statistically evaluate the questions they claim to be evaluating. If the generalizability crisis does indeed come to the forefront – if enough people do as Yarkoni has suggested and stop honoring the charade of statistical inference – none of the major subfields of psychology will be safe. Even so, the costs for clinical psychology should and will be steep.

All of psychology claims to do impactful science. Yet clinical psychology's perennial claims of alleviating suffering and literally saving lives are unique among the subfields of our science. The logical upshot of such claims is that the costs of a generalizability crisis are measured in human lives, not wasted resources. Yarkoni closes by offering readers a choice. The first option is the choice to improve, to make more reasonable inferences, temper one's claims, and return to more basic forms of research such as descriptive and qualitative analyses. The second is to simply stick one's head in the sand and recommit to the cargo cult rituals that have brought success to so many for so long. Although the latter of these two options is certainly the easier path, for psychological scientists in applied domains, particularly those in clinical science, the human costs of ignoring such a crisis must leave them with only one choice.

Conflict of interest. The author declares no conflict of interest in writing this commentary. The author declares no specific funding for this work.

References

Baumeister, R. F., & Vohs, K. D. (2007). Self-regulation, ego depletion, and motivation. *Social and Personality Psychology Compass*, *1*(1), 115–128. https://doi.org/10.1111/j.1751-9004.2007.00001.x.

Berkman, E. T., & Wilson, S. M. (2021). So useful as a good theory? The practicality crisis in (social) psychological theory. *Perspectives on Psychological Science*, *16*(4), 864–874. https://doi.org/10.1177/1745691620969650.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626.

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., … Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*(1), 14–24. https://doi.org/10.1177/1948550619887702.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290.

Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, *37*(7), 10–13.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063.

Friborg, O., Barlaug, D., Martinussen, M., Rosenvinge, J. H., & Hjemdal, O. (2005). Resilience in relation to personality and intelligence. *International Journal of Methods in Psychiatric Research*, *14*(1), 29–42. https://doi.org/10.1002/mpr.15.

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, *23*(2), 107–131. https://doi.org/10.1177/1088868318762183.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873.

Hughes, B. M. (2018). *Psychology in crisis*. Macmillan International Higher Education.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review*, *33*(7), 883–900. https://doi.org/10.1016/j.cpr.2012.09.008.

Lundh, L.-G. (2019). The crisis in psychological science and the need for a person-oriented approach. In J. Valsiner (Ed.), *Social philosophy of science for the social sciences* (pp. 203–223). Springer International Publishing. https://doi.org/10.1007/978-3-030-33099-6_12.

Meehl, P. E. (1987). Theory and practice: Reflections of an academic clinician. In E. F. Bourg, R. J. Bent, J. E. Callan, N. F. Jones, J. McHolland, & G. Stricker (Eds.), *Standards and evaluation in the education and training of professional psychologists: Knowledge, attitudes, and skills* (pp. 7–23). Transcript Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195.

Sakaluk, J. K., Williams, A. J., Kilshaw, R. E., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology*, *128*(6), 500–509. https://doi.org/10.1037/abn0000421.

Schmidt, F. T. C., Nagy, G., Fleckenstein, J., Möller, J., & Retelsdorf, J. (2018). Same same, but different? Relations between facets of conscientiousness and grit. *European Journal of Personality*, *32*(6), 705–720. https://doi.org/10.1002/per.2171.

Stewart, R. E., Stirman, S. W., & Chambless, D. L. (2012). A qualitative investigation of practicing psychologists' attitudes toward research-informed practice: Implications for dissemination strategies. *Professional Psychology: Research and Practice*, *43*(2), 100–109. https://doi.org/10.1037/a0025694.

The role of generalizability in moral and political psychology

Elizabeth A. Harris[a], Philip Pärnamets[b], William J. Brady[c], Claire E. Robertson[a] and Jay J. Van Bavel[a]

[a]Department of Psychology, New York University, New York, NY 10003, USA; [b]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm 7165, Sweden and [c]Department of Psychology, Yale University, New Haven, CT 06250, USA

eah561@nyu.edu; cer493@nyu.edu; jay.vanbavel@nyu.edu; philip.parnamets@ki.se; william.brady@yale.edu

doi:10.1017/S0140525X2100042X, e19

Abstract

The aim of the social and behavioral sciences is to understand human behavior across a wide array of contexts. Our theories often make sweeping claims about human nature, assuming that our ancestors or offspring will be prone to the same biases and preferences. Yet we gloss over the fact that our research is often based in a single temporal context with a limited set of stimuli. Political and moral psychology are domains in which the context and stimuli are likely to matter a great deal (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). In response to Yarkoni (see BBS issue), we delve into topics related to political and moral psychology that likely depend on features of the research. These topics include understanding differences between liberals and conservatives, when people are willing to sacrifice someone to save others, the behavior of political leaders, and the dynamics of intergroup conflict.

The aim of the social and behavioral sciences is to understand human behavior across a wide array of contexts. Our theories often make sweeping claims about human nature, assuming that our ancestors or offspring will be prone to the same biases and preferences. Yet we gloss over the fact that our research is often based on a single temporal context with a limited set of stimuli. Political and moral psychology are domains in which the context and stimuli are likely to matter a great deal (Brandt & Wagemans, 2017; Van Bavel et al., 2016). Understand differences between liberals and conservatives, when people are willing to sacrifice someone to save others, the behavior of political leaders, and the dynamics of intergroup conflict, all likely depend on the features of the research.

One of the major challenges of political psychology is understanding the generalizability of stimuli (see Yarkoni, this issue). Political psychologists often make broad claims about the differences (or lack thereof) between conservatives and liberals. However, the field's findings can vary widely depending on the specific stimuli used (e.g., topics, policies, etc.). Whether American Democrats or Republicans, for example, report higher belief superiority depends on the political issue researchers asked about (Toner, Leary, Asher, & Jongman-Sereno, 2013). Scientists can therefore select different issues to show, alternatively, that people on the left or right of the political spectrum are more prone to a sense of superiority about their beliefs. It is only when looking across multiple topics that the overall quadratic relationship emerges (Harris & Van Bavel, 2021), suggesting the people on both ends of the spectrum are similarly prone to feelings of superiority when their attitudes are extreme.

This issue of non-generalizable stimuli extends to research on moral psychology as well. For instance, the victims of harm are often described using vague terminology, such as five nondescript workers on a track about to be run over by a trolley. In many studies, the victims are genderless, raceless, etc. (Hester & Gray, 2020), which makes it difficult to generalize the findings to more vivid or real-world judgments (FeldmanHall et al., 2012). Yet the research suggests that moral judgments are heavily influenced by contextual information and stimuli.

Another often-overlooked aspect of generalizability in both political and moral psychology is the changing temporal dynamics. Historical evidence demonstrates that core concepts studied by political psychologists, such as partisanship and polarization, are changing across time (Abramowitz & Webster, 2016;

Baldassarri & Gelman, 2008; Bartels, 2000; Kozlowski & Murphy, 2020). Furthermore, the stimuli used to study these phenomena often include specific policies or politicians who are well known at a given time and elicit a host of associations related to that period of time.

Temporal factors occurring "outside the lab" may also influence the interactions between moral and political psychology. For instance, "political elites" in the United States use more moral language when their party is not in power (Wang & Inbar, 2020), demonstrating the changing nature of moral language usage due to external (and typically unacknowledged) factors (see Brady, Wills, Burkart, Jost, & Van Bavel, 2019). Similarly, moral language changes over time (cf. Wheeler, McGrath, & Haslam, 2019). The question of temporal generalizability is a foundational question of the historical versus lawlike nature of psychological findings (Gergen, 1973). A critical next step in psychological research must be examining whether trends related to political or moral reasoning are robust across time.

It is also important to avoid the assumption that reactions to political and moral stimuli in one cultural context reveal insights into individuals in another cultural context. For example, claims about growing rates of political polarization are often made based on US samples, with participants embedded within a two-party political system (Finkel et al., 2020). Recent evidence suggests that other countries show varying trajectories of political polarization likely based on context-specific characteristics of their political system (Boxell, Gentzkow, & Shapiro, 2020). Future work should include international samples or else constrain their claims to reflect the current sample.

In moral psychology, foundational research on moral intuitions largely ignored the prospect of cultural variation. For example, dual-processing theories developed by studying trolley problem dilemmas exclusively used evidence from WEIRD [western, educated, industrialized, rich, and democratic] samples (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), yet recent cross-cultural work demonstrates that participants' intuitions about trolly problem dilemmas are notably different in eastern cultures (Rehman & Dzionek-Kozłowska, 2020) and non-industrialized cultures (Sorokowski, Marczak, Misiak, & Białek, 2020). Moral and political psychologists should be aware of the type of evidence required to make claims of strong or weak universality (Norenzayan & Heine, 2005), and make explicit reference to cultural generalizability when drawing broad conclusions from a small number of studies or restricted samples (Bago et al., 2021).

Moral and political psychology face major generalizability challenges with regard to the content of research materials, temporal variability, and cultural differences. To address these challenges, part of the solution involves methodological and analytic changes (target article by Yarkoni). While important, these changes alone cannot fully address questions of generalizability. Here we join many recent observers in calling for better theory development and employment of formal tools to make more precise theoretical claims (Guest & Martin, 2020; Muthukrishna & Henrich, 2019; van Rooij & Baggio, 2021). Improved theory will not only help close the distance between theoretical claims and empirical tests, but guide and facilitate interdisciplinary research. The latter may be particularly important in moral and political psychology, as its subject matter clearly intersects with explanatory efforts in many adjacent disciplines.

One hope is that, with changes to our empirical and theoretical practices, future researchers will be better placed to assess the generalizability of research in moral and political psychology. One potential future for this work is for researchers to organize themselves not only in cross-disciplinary and cross-cultural consortia but also cross-temporally. This would mean planning for research projects to investigate fundamental questions over longer time periods that span entire grant cycles or even individual careers. Conducting research with more diverse samples of participants and stimuli, across social and political contexts, will provide an exciting foundation for the future of the field.

Financial support. This publication was made possible through the support of: John Templeton Foundation (#61378; PP and JVB), Swedish Research Council (2016-06793; PP), and the Social Science and Humanities Research Council of Canada (752-2018-0213; EH).

Conflict of interest. None.

References

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of US elections in the 21st century. *Electoral Studies*, *41*, 12–22. https://doi.org/10.1016/J.ELECTSTUD.2015.11.001.

Bago, B., Aczel, B., Kekecs, Z., Protzko, J., Kovacs, M., Nagy, T., … Dutra, N. B. (2021). Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgments. https://doi.org/10.31234/osf.io/9uaqm.

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408–446. https://dx.doi.org/10.2139%2Fssrn.1010098.

Bartels, L. M. (2000). Partisanship and voting behavior, 1952–1996. *American Journal of Political Science*, *44*(1), 35–50.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2020). *Cross-country trends in affective polarization* (No. w26669). National Bureau of Economic Research. http://www.nber.org/papers/w26669.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*(10), 1802–1813. https://doi.org/10.1037/xge0000532.

Brandt, M. J., & Wagemans, F. (2017). From the political here and now to generalizable knowledge. *Translational Issues in Psychological Science*, *3*(3), 317–320. https://doi.org/10.1037/tps0000126.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. http://dx.doi.org/10.1016/j.cognition.2012.02.001.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., IIyengar, S., Klar, S., … Druckman, J. N. (2020). Political sectarianism in America: A poisonous cocktail of othering, aversion, and moralization. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/SCIENCE.ABE1715.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309–320. https://doi.org/10.1037/h0034436.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872.

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. https://doi.org/10.31234/osf.io/rybh9.

Harris, E. A., & Van Bavel, J. J. (2021). Preregistered replication of "feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority." *Psychological Science*, *32*(3), 451–458. https://doi.org/10.1177/0956797620968792.

Hester, N., & Gray, K. (2020). The moral psychology of raceless, genderless strangers. *Perspectives on Psychological Science*, *15*(2), 216–230. https://doi.org/10.1177%2F1745691619885840.

Kozlowski, A. C., & Murphy, J. P. (2020). Issue alignment and partisanship in the American public: Revisiting the 'partisans without constraint' thesis. *Social Science Research*, *94*, Article 102498.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, *131*(5), 763–784. https://psycnet.apa.org/doi/10.1037/0033-2909.131.5.763.

Rehman, S., & Dzionek-Kozłowska, J. (2020). The Chinese and American students and the trolley problem: A cross-cultural study. *Journal of Intercultural Communication*, *20*(2), 31–41.

Sorokowski, P., Marczak, M., Misiak, M., & Białek, M. (2020). Trolley dilemma in Papua. Yali horticulturalists refuse to pull the lever. *Psychonomic Bulletin & Review*, *27*, 398–403. https://doi.org/10.3758/s13423-019-01700-y.

Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, *24*(12), 2454–2462. https://doi.org/10.1177/0956797613494848.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177%2F1745691620970604.

Wang, S. Y. N., & Inbar, Y. (2020). Moral-language use by US political elites. *Psychological Science*, *32*(1), 14–26. https://doi.org/10.1177%2F0956797620960397.

Wheeler, M. A., McGrath, M. J., & Haslam, N. (2019). Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. *PLoS ONE*, *14*(2), e0212267. https://doi.org/10.1371/journal.pone.0212267.

Without more theory, psychology will be a headless rider

Witold M. Hensel[a], Marcin Miłkowski[b] and Przemysław Nowakowski[b]

[a]Institute of Philosophy, University of Bialystok, pl. NZS 1, 15-420 Białystok, Poland and [b]Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland.
whensel@poczta.onet.pl
mmilkows@ifispan.edu.pl
pnowakowski@ifispan.edu.pl; http://marcinmilkowski.pl/

doi:10.1017/S0140525X21000212, e20

Abstract

We argue that Yarkoni's proposed solutions to the generalizability crisis are half-measures because he does not recognize that the crisis arises from investigators' underappreciation of the roles of theory in experimental research. Rather than embracing qualitative analysis, the research community should make an effort to develop better theories and work toward consistently incorporating theoretical results into experimental practice.

Yarkoni presents the psychologist with a choice: either embrace qualitative analysis or else adopt the specific solutions described in section 6.3 and suffer the consequences. Yet the initial choice itself is a false dilemma and the solutions are half-measures. The rub is that Yarkoni's proposals are based on three dubious assumptions: (1) that empirical science is only about collecting and analyzing data, which leaves theoretical investigation almost completely out of the picture, (2) that the distinction between quantitative and qualitative research is fundamental to addressing the crisis, and (3) that qualitative research is, essentially, quantitative inquiry *sans* inferential statistics.

As to assumption (3), qualitative investigation differs from quantitative research in many important respects, determined in the final analysis by the different aims the two kinds of inquiry are intended to achieve. Qualitative researchers rely on in-depth interviews, observation, and document analysis rather than experimentation. Also, the logic behind purposeful sampling used in qualitative research is, roughly speaking, a mirror image of the logic behind statistical sampling: what is a strength in one would be a weakness in the other (Patton, 2005). Therefore, Yarkoni is wrong when he says that, in many subfields of psychological science, embracing qualitative analysis would amount to merely dropping statistical inference. That would lead to replacing one kind of poor quality research by another, and doing poor quality research is no answer to any crisis. Correspondingly, while we have nothing against good qualitative inquiry, abandoning quantitative research in favor of it would amount to throwing the baby out with the bathwater. The two approaches are not mutually exclusive but complementary (Shadish, 1993). Psychology needs both.

So, to address the generalizability crisis, we must fix the way quantitative research is done. The question is: how? Here, Yarkoni's assumptions (1) and (2) intervene by distorting and restricting available options. Interestingly, although assumption (1) blinds Yarkoni to the significance of theoretical work in general, he wouldn't have been able to make his case without appealing to theoretical insights. He resolves the tension by conflating theorizing with qualitative analysis. This is a mistake. Theorizing is an activity integral to any scientific approach regardless of its specific aims and methods. It transcends the difference between qualitative and quantitative research. This, we believe, is crucial because all the shortcomings of current practice discussed by Yarkoni come from a common source: researchers' inadequate appreciation of how various theoretical considerations should inform the decisions made at every stage of scientific investigation.

Consider Yarkoni's critique of Alogna et al.'s (2014) many-lab replication of verbal overshadowing. From our perspective, it showcases that choices regarding which phenomena to study empirically, or which results to replicate, ought to be made against a broader theoretical background that includes general theoretical principles such as "the most basic, uncontroversial facts about the human mind" cited by Yarkoni. Ideally, such principles should be systematized to provide insights into the human mind, forming a theoretical framework to guide further research (Irvine, 2021; Muthukrishna & Henrich, 2019). But they should not be ignored even in the absence of such a framework.

The generalizability crisis, by contrast, is more to do with local theories. It arises from widespread failures to appreciate the relations between the rich conceptual variables employed by micro- or middle-range theories (Cartwright, 2020), on one hand, and their operationalizations, on the other (cf. Shadish, Cook, & Campbell, 2002). These failures often yield inconsistencies between data and various components of accepted theory, understood as a coherent and well-ordered set of concepts or models. While Yarkoni rightly draws attention to the mismatch between verbal and mathematical descriptions, in many cases, the inconsistencies can be understood without special statistical training – for example, the nature and possible ramifications of mono-operation and mono-method biases are easy to comprehend (Shadish et al., 2002, 75–76).

To tie Yarkoni's critique of Alogna et al. (2014) with the generalizability crisis, note that inattention to theory can affect generalizability even if some of the resulting issues are not directly related to operationalization. For example, attempts to integrate research associated with mutually inconsistent theoretical frameworks can cause confusion and thereby affect validity across the board – as has recently been observed in research on emotion (Weidman, Steckler, & Tracy, 2017).

The moral we draw is that the generalizability crisis is unlikely to go away as long as the community as a whole does not recognize that an adequate understanding of theory is essential to research validity. It is theory, not gut feelings, that should tell us what kind of data to collect, how to collect them, and how

to analyze and interpret them (Kukla, 1989). Furthermore, *pace* Yarkoni, theory cannot be read off of empirical data: theory needs to be *developed*, which requires a set of skills different from that of the experimenter. Like many other sciences, psychology needs specialized theorists whose work visibly contributes to experimental research (MacKay, 1988).

Let us close by calling attention to important similarities between the generalizability crisis and the replication crisis. Both have been with us for quite some time and both involve widespread violation of fundamental and well-known principles of scientific investigation. It is fairly obvious, for example, that the findings of a single small study may very well be false positives, especially after some p-value hacking. It is equally obvious that the inferences we draw from obtained data should be warranted. Arguably, researchers do not need Yarkoni to educate them about the need for conservative conclusions: they know the rules – they just do not follow them. This suggests that we should explore measures focused on changing the research culture (Nosek, Spies, & Motyl, 2012). But although many practices advocated by the open science movement, such as data sharing and improved quality of reporting (Hensel, 2020; Miłkowski, Hensel, & Hohol, 2018), can help to enhance both reproducibility and generalizability (the latter, by enabling high-quality re- and meta-analysis), it is also necessary to strengthen theorizing and work toward consistently incorporating theoretical results into experimental research. Without that, psychology will be a headless rider doomed to face ever new crises.

Financial support. This work was supported by the National Science Centre (Poland) research grant (MM, grant number 2014/14/E/HS1/00803).

Conflict of interest. None.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653.

Cartwright, N. (2020). Middle-range theory: Without it what could anyone do?. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *35*(3), 269–323.https://doi.org/10.1387/theoria.21479.

Hensel, W. M. (2020). Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience. *European Journal for Philosophy of Science*, *10*, 44. https://doi.org/10.1007/s13194-020-00317-6.

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, *16*(4), 844–853. https://doi.org/10.1177/1745691620970558.

Kukla, A. (1989). Nonempirical issues in psychology. *American Psychologist*, *44*(5), 785–794. https://doi.org/10.1037/0003-066X.44.5.785.

MacKay, D. G. (1988). Under what conditions can theoretical psychology survive and prosper? Integrating the rational and empirical epistemologies. *Psychological Review*, *95*(4), 559–565. https://doi.org/10.1037/0033-295X.95.4.559.

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, *45*(3), 163–172. https://doi.org/10.1007/s10827-018-0702-z.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058.

Patton, M. Q. (2005). Qualitative research. In Everitt, B. S., & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1633–1636). John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa514.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *60*, 13–57. https://doi.org/10.1002/ev.1660.

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Houghton-Mifflin.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/emo0000226.

Citizen science can help to alleviate the generalizability crisis

Courtney B. Hilton[a] and Samuel A. Mehr[a,b,c]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138, USA; [b]Data Science Initiative, Harvard University, Cambridge, MA 02138, USA and [c]School of Psychology, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand.
courtneyhilton@g.harvard.edu; sam@wjh.harvard.edu; https://themusiclab.org

doi:10.1017/S0140525X21000352, e21

Abstract

Improving generalization in psychology will require more expansive data collection to fuel more expansive statistical models, beyond the scale of traditional lab research. We argue that citizen science is uniquely positioned to scale up data collection and, that in spite of certain limitations, can help to alleviate the generalizability crisis.

Yarkoni argues that common statistical practices in psychology fail to quantitatively support the generalizations psychologists care about. This is because most analyses ignore important sources of variation and, as a result, unjustifiably generalize from narrowly sampled particulars.

Is this problem tractable? We are optimists, so we leave aside Yarkoni's suggestions to "do something else" or "embrace qualitative research," and focus instead on his key prescription: the adoption of mixed-effects modeling to estimate effects at the level of a factor (e.g., stimulus), to be interpreted as one of a population of potential measurements, licensing generalization over that factor.

Yarkoni is correct that far too few studies do this. In our field of the psychology of music, many inaccurately generalize, for example, from a single musical example to *all* music; or from a set of songs from a particular context (e.g., pop songs) to *all* songs; or from the music perception abilities of a particular subset of humans to *all* humans.

Consider the "Mozart effect": a notorious positive effect of listening to a Mozart sonata on spatial reasoning that was overgeneralized to "all Mozart" and eventually "all music." While replicable under narrow conditions, the original result was, in fact, specific to neither spatial reasoning, Mozart, nor music generally – the effect was the result of generic modifications to arousal and mood (Thompson, Schellenberg, & Husain, 2001).

Modeling random effects for stimuli and other relevant factors, however, brings with it a substantial challenge: researchers will need far more stimuli and participants, sampled more broadly and deeply, and with far more measures, than is typically practical. Psychologists already struggle to obtain sufficient statistical power for narrowly sampled, fixed-effect designs (Smaldino & McElreath, 2016).

How, then, can we alleviate the generalizability crisis? We think *citizen science* can help.

Citizen science refers to a collection of research tools and practices united by the alignment of interests between participants and the aims of the project, such that participation is intrinsically motivated (e.g., by curiosity in the topic) rather than by extrinsic factors (e.g., money or course credit). The results are studies that cheaply recruit thousands or even millions of diverse participants via the internet. Studies take many forms, ranging from "gamified" experiments that go viral online, such as our "Tone-deafness test" (current $N > 1.2$ million; https://themusiclab.org); to collective/collaborative field reporting, such as New Zealand's nationwide pigeon census (the Great Kererū count, https://www.greatkererucount.nz/).

The potential of citizen science is staggering. For example, the Moral Machine Experiment (Awad et al., 2018) collected 40 million decisions from millions of people (representing 10 languages and over 200 countries) on moral intuitions about self-driving cars. Such massive scale enabled the quantification of cross-country variability in moral intuitions, and how it was mediated by cultural and economic factors particular to each country, with profound real-world implications.

Further, when citizen science is coupled with corpus methods, generalizability across stimuli can be effectively maximized. We previously investigated high-level representations formed during music listening, by asking whether naïve listeners can infer the behavioral context of songs produced in unfamiliar foreign societies (Mehr et al., 2018, 2019). Each iteration of a viral "World Music Quiz" played a random draw of songs from the *Natural History of Song* corpus, a larger stimulus set that representatively samples music from 86 world cultures.

As such, the findings of the experiment – that listeners made accurate inferences about the songs' behavioral contexts – can be accurately generalized (a) to the populations of songs the stimulus subsets were drawn from (e.g., lullabies); (b) more weakly, to music, writ large (insofar as the subpopulations of songs represented by those categories sample from other categories); and (c) to the population of listeners from whom our participants were drawn (i.e., members of internet-connected societies). All of these factors can be explicitly modeled with random effects.

The same reasoning applies to studying subpopulations of participants (measured in terms of any characteristic) and even subsets of corpora. For example, in a study of acoustic regularities in infant-directed vocalizations across cultures, we model random effects of listener characteristics, speaker/singer (i.e., the producers of the stimuli) characteristics, and stimulus categories of interest (e.g., infant-directed vs. adult-directed speech). This is only possible with large datasets (in our case, nearly 1 million listener judgements; Hilton, Moser, et al., 2021). Other under-used analyses also become more practical with big citizen-science data, including radical randomization (Baribault et al., 2018), prediction with cross-validation (Yarkoni & Westfall, 2017), and matching methods for causal inference (Stuart, 2010).

Citizen-science methods are limited, however, by the need to factor in participants' interests and incentives; the need to avoid factors that might dissuade participation (e.g., clunky user interfaces, boring time-consuming tasks), which can require graphic design and web development talent for "gamification" (e.g., Cooper et al., 2010); the risks of recruiting a biased population subset (i.e., those with internet access; Lourenco & Tasimi, 2020); and the trade-offs between densely sampling stimuli across- versus within-participants, given the typically short duration of citizen-science experiments.

Indeed, while our efforts to recruit children at scale online via citizen science show promising results (Hilton, Crowley de-Thierry, Yan, Martin, & Mehr, 2021), rare or hard-to-study populations may be difficult to recruit en masse (cf. Lookit, a platform for online research in infants; Scott & Schulz, 2017). As Yarkoni notes, alternative approaches like multisite collaborations (e.g., ManyBabies Consortium, 2020) could be calibrated to maximize generalizability across stimuli rather than directly replicating results with the same stimuli.

All that being said, thanks to a growing ecosystem of open-source tools (e.g., de Leeuw, 2015; Hartshorne, de Leeuw, Goodman, Jennings, & O'Donnell, 2019; Peirce et al., 2019); the availability of large-scale, naturalistic corpora from industry partners (e.g., Spotify Research; Way, Garcia-Gathright, & Cramerr, 2020); and calls for collaborative, field-wide investment in citizen-science infrastructure (Sheskin et al., 2020) – addressing these limitations has never been easier.

As such, we think that citizen science can play a useful role as psychologists begin to address the generalizability crisis.

Acknowledgment. We would like to thank Max Krasnow, Mila Bertolo, Stats Atwood, Alex Holcombe, and William Ngiam for feedback on drafts of this commentary.

Financial support. C.B.H. and S.A.M. are supported by NIH DP5OD024566. S.A.M. is supported by the Harvard Data Science Initiative.

Conflict of interest. None.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Sharff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., … Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Hartshorne, J. K., de Leeuw, J., Goodman, N., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1–22.

Hilton, C., Crowley de-Thierry, L., Yan, R., Martin, A., & Mehr, S. (2021). Children infer the behavioral contexts of unfamiliar songs. *PsyArXiv*. doi: 10.31234/osf.io/rz6qn.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., … Mehr, S. A. (2021). Acoustic regularities in infant-directed vocalizations across cultures. *bioRxiv*. doi: 10.1101/2020.04.09.032995

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*, 24–52.

Mehr, S. A., Singh, M., Knox, D., Ketter, D., Pickens-Jones, D., Atwood, S., … Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868.

Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, *28*(3), 356–368.e5.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., … Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science*, *12*(3), 248–251.

Way, S. F., Garcia-Gathright, J., & Cramerr, H. (2020). Local trends in global music streaming. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 10.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Look to the field

Rumen Iliev[a], Douglas Medin[b] and Megan Bang[c]

[a]Toyota Research Institute, Los Altos, CA 94022, USA; [b]Department of Psychology and School of Education and Social Policy, Northwestern University, Evanston, IL 60208, USA and [c]Learning Sciences and Department of Psychology, Northwestern University, Annenberg Hall, Evanston IL 60208, USA
rumen.iliev@tri.global
medin@northwestern.edu
Megan.bang@northwestern.edu

doi:10.1017/S0140525X21000509, e22

Abstract

Yarkoni's paper makes an important contribution to psychological research by its insightful analysis of generalizability. We suggest, however, that broadening research practices to include field research and the correlated use of both converging and complementary observations gives reason for optimism.

We agree with Yarkoni's thesis that there is a "generalizability crisis" and that the mapping between verbal theoretical constructs and measures and models is the source of many difficulties. In particular, the limited variation in procedures, stimuli, contexts, and measures represents a significant challenge to generalizability. Yarkoni summarizes these concerns by suggesting that "a huge proportion of the quantitative inferences drawn in the published psychology literature are so weak as to be at best questionable and at worst utterly nonsensical."

Although Yarkoni's arguments are compelling, we don't fully agree with the somewhat gloomy picture he paints. The generalizability crisis creates something of a paradox: If generalization claims are on such shaky grounds, why is it that many phenomena are so robust that they make for reliable classroom demonstrations and/or have been shown to have substantial practical significance?

With respect to the former, examples include a number of judgment and decision biases identified and analyzed by Kahneman, Tversky, Fischhoff, Slovic, Loewenstein, Weber, and others (e.g., availability heuristic, loss aversion, framing effects, quantity insensitivity). With respect to the latter, Cialdini (2009a, 2009b) has demonstrated simple but effective manipulations that increase environmentally friendly behaviors (e.g., hotel guests reusing towels). Similarly, implementing changes default assumptions (Thaler & Sunstein, 2008) has been shown to facilitate policy goals such as increasing organ donation.

Field versus lab

We suggest that attention to the field is a critical factor supporting both relevance and generalizability. Those involved in lab research usually aim to demonstrate the presence of a particular effect, and tend to be motivated to create a specific environment or context to observe it. Lab researchers have an unlimited number of levers to establish conditions which will maximize the chances for observing desired effects. Rigorous control procedures can be implemented that are not feasible outside the lab. But this precise control may be exactly what limits generalizability.

Field researchers face the opposite problem. They typically work in environments which can be changed very little, and with populations they rarely can preselect. Field/applied researchers are routinely motivated to search for effects and manipulations which are robust enough to work in their specific context. Field research may operate as a "generalizability filter" separating tenuous effects from interventions with a higher chance for success.

Judgment and decision-making research may have benefited from the fact that much of it has been done in business schools. Business school faculty rarely have access to a "subject pool" and they tend to rely on both studies in classrooms and in the field. The participants in business school studies often are students who have experience in the business world and are seeking MBAs (or PhDs). This is just one factor that serves to increase the likelihood that research by business school faculty will make connections with corporate contexts.

Consider, for example, "sunk cost" effects. Sunk costs refer to situations where commitment of resources is continued and escalated beyond any rational considerations because one doesn't want to "waste" the prior investment. This is sometimes referred to as "throwing good money after bad." The interest in sunk cost effects originated with real-world examples. But a careful analysis of generalizability suggests that there are other situations where the opposite of sunk cost effects can be shown (prematurely withdrawing an investment just before it starts to pay off; e.g., Drummond, 2014; Heath, 1995). Instead of undermining the sunk costs construct, such findings invite attention to what factors are associated with each type of outcome. For instance, sunk cost effects for money may be different from sunk cost effects for time (Cunha, Marcus, & Caldieraro, 2009; Soman, 2001).

Field research may also serve as a direct test of generalizability of lab findings. For example Hofmann, Wisneski, Brant, and Skitka (2014) used text messaging at varied times to assess everyday moral and immoral acts and experiences. They found moral experiences to be common and, they observed both moral licensing and moral contagion, effects that previously had been shown in lab studies.

This interplay between lab and field is useful to both. Although generalizability is important, it could be argued that variability is even more fundamental. At the heart of social science is the search for patterned variation, variation that our theories seek to understand. Attention to the field may serve to increase attention to potential interactions and undermine a main effect focus.

Field as a source of complementary evidence

As Yarkoni notes, conceptual replications (as opposed to exact replications) put assumptions of generalizability to the test and represent an effective research strategy. They also are a key tool in establishing construct validity (e.g., Grahek, Schaller, & Tackett, 2021), linking theory and measures.

Field observation offers a complementary form of converging measure that can be an important research tool. For example, lab studies suggesting that participants see nature as incompatible with human presence (nature is pristine and humans can enjoy it but are not part of it) can be complemented by analyses using Google images. For example, a search of images for "ecosystems" found that humans were present only two percent of the

time and for about half of that two percent humans were outside the system looking in (Medin & Bang, 2014). Similarly, experimental observations suggesting cultural differences in subjective proximity to nature (Bang, Medin, & Atran, 2007) may be complemented by corresponding differences in illustrations in children's books (Bang et al., in press).

An additional benefit of complementary field observations is that they facilitate analyzing changes over time (Iliev & Ojalehto, 2015). For example, claims about increasing cultural individualism may be paralleled by corresponding changes in cultural artifacts (Greenfield, 2013, 2017). In short, field observations invite complementary and coordinated observations both as a stimulus for new studies and as a guide to robustness of findings.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Bang, M., Alfonso, J., Faber, L., Marin, A., Marin, M., Medin, D., … Woodring, J. (in press). Perspective taking and psychological distance in children's picture books: Differences between native and non-native authored books. In S. Nelson-Barber & P. W. U. Chinn (Eds.), *Indigenous STEM education: Perspectives from the Pacific Islands, the Americas and Asia*. New York, NY: Springer.

Bang, M., Medin, D. L., & Atran, S. (2007). Cultural mosaics and mental models of nature. *Proceedings of the National Academy of Sciences*, *104*(35), 13868–13874.

Cialdini, R. B. (2009a). *Influence: Science and practice* (5th ed.). Boston: Allyn & Bacon.

Cialdini, R. B. (2009b). We have to break up. *Perspectives on Psychological Science*, *4*, 5–6.

Cunha, Jr. M., & Caldieraro, F. (2009). Sunk-cost effects on purely behavioral investments. *Cognitive Science*, *33*(1), 105–113.

Drummond, H. (2014). Escalation of commitment: When to stay the course? *Academy of Management Perspectives*, *28*(4), 430–446.

Grahek, I, Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives in Psychological Science*, *16*(4), 803–815.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science*, *24*, 1722–1731.

Greenfield, P. M. (2017). Cultural change over time: Why replicability should not be the gold standard in psychological science. *Perspectives on Psychological Science*, *12*(5), 762–771.

Heath, C. (1995). Escalation and de-escalation of commitment in response to sunk costs: The role of budgeting in mental accounting. *Organizational Behavior and Human Decision Processes*, *62*, 38–54.

Hofmann, W., Wisneski, D. C., Brant, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*, 1340–1343.

Iliev, R. L., & Ojalehto, B. L. (2015). Bringing history back to culture: On the missing diachronic component in the research on culture and cognition. *Frontiers in Psychology 10*, 1–4.

Medin, D. L., & Bang, M. (2014). *Who's asking? Native science, western science and science education*. MIT Press.

Soman, D. (2001). The mental accounting of sunk time costs: Why time is not like money. *Journal of Behavioral Decision Making*, *14*(3), 169–185.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin Books.

Science with or without statistics: Discover-generalize-replicate? Discover-replicate-generalize?

John P.A. Ioannidis

Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA.
jioannid@stanford.edu

doi:10.1017/S0140525X21000054, e23

Abstract

Overstated generalizability (external validity) is common in research. It may coexist with inflation of the magnitude and statistical support for effects and dismissal of internal validity problems. Generalizability may be secured before attempting replication of proposed discoveries or replication may precede efforts to generalize. These opposite approaches may decrease or increase, respectively, the use of inferential statistics with advantages and disadvantages.

Inflated claims are prevalent in research and the reward system facilitates them (Smaldino & McElreath, 2016). Both the magnitude and statistical support of effects, but also the narratives researchers craft of these effects, can be inflated. To evaluate effect inflation, one can scrutinize numbers presented with specific metrics and/or subject to specific statistical inferential tools. Errors and deficiencies in internal validity can also be modeled or probed within the same quantitative machinery. Conversely, inflation in the accompanying narrative that tries to instill meaning, relevance, and breadth into scientific investigation evades quantification. Some of that inflation pertains to silencing or underestimating internal validity problems and limitations. The most egregious narrative boosting, however, pertains to external validity, aka generalizability. Ignoring, silencing, or downplaying sources of variability; putting a spin to read the results as more important than they are (Boutron & Ravaud, 2018); extrapolating to a broader paradigm than narrowly focused data would allow – are all common problems. Moreover, for applied research that carries decision-making implications, inferring broadly actionable results is the end-product of that expansive narrative.

Inflation of effects, downplaying of internal validity concerns and overstated generalizability often coexist. It is easier to overstate generalizability when effect sizes, statistical significance, or any other type of statistical support seem stronger, thus more immune to error. Supposedly, stronger effects may withstand a greater assault from bias and allow a greater leap of faith for their generalizability. However, this is a misconception. In reality, the opposite may be true. Large effects and strong statistical support may simply herald the presence of more bias and least generalizability, that is, deficits in internal or external validity or both (Ioannidis, 2016). The most erroneous data and studies and the more extreme, outlying, non-representative situations and conditions may yield the most astonishing large effects. Whenever scientists come upon discovering a large effect in their research endeavors, they should be particularly worried. The first step should be to go back and find out where some major error has occurred. When no error is found, the second step is to think why this stupendous effect may represent a very unusual situation, with little or no relevance in most other settings.

In trying to remedy this situation, different solutions have been proposed and some of them are pulling in opposite directions. To neutralize excessive, unwarranted claims of discovered effects, one solution is to submit them to exact replication with the hope that, if properly done, false-positive effects should be refuted (Nosek & Errington, 2020). The sequence goes: discover-replicate-generalize, or, in other words, try to replicate first and, if it replicates, then try to see how far the research finding can generalize to other, different, expansive settings. A second

solution, espoused by Yarkoni, is to give priority to generalizability. The sequence goes: discover-generalize-replicate, that is, don't waste time with replication unless a promising research finding has been probed in a sufficiently large variety of settings to have some sense that it is generalizable (and even remotely worthy). In the extreme form, this approach would give the search for generalizability not just priority but also dominance. Research would be mostly an exploration of variability and of the boundaries of generalizability.

These solutions may have different implications for the extent to which inferential statistics should be used. The "discover-replicate-generalize" sequence would require inferential statistics to be deployed, and strengthened, if anything, compared with current practices. Other safeguards such as prespecification and registration are also essential. Not only the main effects, but also issues of their internal validity should be modeled as rigorously as possible with the best statistical methods and inference tools. In fact, if internal validity cannot be secured or taken properly into account with some proper quantitative methods, rushing into replication would be a nuisance: the same errors will be carried forward unopposed and unaccounted for.

Conversely, with the "discover-generalize-replicate" sequence, it is tempting to postpone and thus diminish the use of inferential statistics in the research process. Research becomes mostly a process of description, a collection of notes and observations, like collecting stamps or butterflies and marveling at how different they are. One may even suspect an undertone of cynicism in this approach: because most observations are likely to be misleading and/or non-generalizable, we should not make too much of them. We should not take them or us, as researchers, too seriously. This guidance aims to avoid having too many false-positives; not by eliminating them, but by not allowing them to be called "positives" in the first place.

The choice between the two strategies is not straightforward – and any choice may not be generalizable! Different disciplines and types of scientific investigation may need a different mix. However, any effort to fix the misuse of statistics simply by removing statistics or statistical rules (no matter how imperfect these rules) may not necessarily make things better and may lead to an even worse "free lunch" situation (Ioannidis, 2019). Weird, exaggerated claims will still be made. In the absence of any statistical obstacle, they may be made even more easily and with even less restrain. At the extreme, the "premium generalize" strategy may end up making science not much different than a competition of fiction writers coming up with qualitative narratives and without any clear rules on what narrative should be preferred over others. For applied science where decisions are pressing, decision-making may become even more subjective and biased – and it is already too subjective and biased in many circumstances.

At the same time, the major problem of over-generalizing with the blessing of statistic rituals, replication, and all cannot be overstated. Poorly used statistics only exacerbate the problem as they give to these misleading claims a false aura of quantitative legitimacy. Perhaps, instead of less statistics and less quantification, one needs more and better. More appropriate models may incorporate more of the known and unknown variability and generate wider (or at least more fair) estimates of uncertainty. Then, perhaps there will be fewer candidates that are considered worthwhile to spend replication efforts – let alone, dare generalize.

Financial support. METRICS has been supported by grants from the Laura and John Arnold Foundation. The work of John Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell.

Conflicts of interest. None.

References

Boutron, I., & Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences of the USA*, *115*(11), 2613–2619.

Ioannidis, J. P. (2016). Exposure-wide epidemiology: Revisiting Bradford Hill. *Statistics in Medicine*, *35*(11), 1749–1762.

Ioannidis, J. P. (2019). The importance of predefined rules and prespecified statistical analyses: Do not abandon significance. *JAMA*, *321*(21), 2067–2068.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

A crisis of generalizability or a crisis of constructs?

Kevin M. King[a] and Aidan G.C. Wright[b]

[a]Department of Psychology, University of Washington, Seattle, WA 98195, USA and [b]Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA

kingkm@uw.edu
http://faculty.washington.edu/kingkm
aidan@pitt.edu
http://www.personalityprocesses.com/

doi:10.1017/S0140525X21000443, e24

Abstract

Psychologists wish to identify and study the mechanisms and implications of nomothetic constructs that reveal truths about human nature and span across operationalizations. To achieve this goal, psychologists should spend more time carefully describing and measuring constructs across a wide range of methods and measures, and less time rushing to explain and predict.

"I live in a jingle jangle jungle. If you ain't got it, you can't be it"
-Bobby Darin

Yarkoni raises concerns about business as usual in psychological science, noting that our methods are rarely designed to extrapolate much beyond the specific sample, measures, or procedures at hand. He aptly frames this as a crisis of generalizability, because if you change samples, measures, or procedures and the results don't hold, then what can be extrapolated? We see at least one alternative way to construe these same issues: namely, a crisis of construct validity. We see this as a valuable alternative articulation, because although many scientists may be willing to write off external validity, knowing that these issues also cut to the core of internal validity may well give them pause for thought.

We argue that most psychologists want to identify and study the mechanisms and implications of nomothetic constructs that

reveal fundamental truths about human nature and expand beyond any specific operationalization. Clinicians want to understand depression, not the Beck Depression Inventory. Personality psychologists want to understand narcissism, not the Narcissistic Personality Inventory. Cognitive psychologists want to understand attention, not the dot probe task. However, to the extent that our methods are too tightly tethered to single methods or measures, we have not elucidated the conceptual, but rather echoed the operational. We risk becoming a science of squares, not circles, in structural equation modeling terms. The key point is that this isn't just a matter of external validity, but it cuts to the core of internal validity, and what it is we think we are studying.

Because the field gives such short shrift to the development of measures that can flexibly, reliably, and broadly capture constructs of interest, the field is polluted with methods and measures that have wide acceptance but perform poorly on some dimension of internal validity. "Gold-standard" measures are only thus because of a field-wide consensus weighed in citations rather than empirical quality. Entire bodies of literature are developed with measures whose psychometric properties have barely been questioned, much less deeply interrogated.

For instance, in some fields, construct validity is largely limited to a reliance on face validity to support construct representation (Whitely, 1983). For example, the ego depletion literature manipulated and measured behaviors ranging from stating the actual color of a color word instead of reading the color word (e.g., a Stroop task), eat healthy foods, regulate emotions, give counter-attitudinal speeches, behave counter to a learned habit, regulate attention, make decisions, or persist in an unpleasant task (Hagger et al., 2016; Hagger, Wood, Stiff, & Chatzisarantis, 2010). However, no research in this domain focused on the fundamental measurement properties of these varying operationalizations. From a construct validity perspective (e.g., Borsboom, Mellenbergh, & Van Heerden, 2004), did they represent some common construct, and can they reliably capture variance attributable to the construct? Although the variety of operationalizations of self-control was admirable, no effort was made to stop and ask whether they reflected the same construct. Researchers would benefit from spending more time simply describing constructs and seeking to sample items and stimuli that might sample as *broad of a range of the construct as possible*, in order to define the limits of what is and what is not a reasonable measure of the construct. As Yarkoni argues, sampling from a broad range of stimuli for both IVs and DVs is a critical method to establish the construct validity.

Another example of how easy it is to put cart before horse when seeking to establish broadly defined constructs is the NIMH Research Domain Criteria (RDoC) (Insel, 2014), which has spent over one billion dollars (per NIH RePORTER) pursuing evidence for neural circuits that span "units" of analysis. The point is not that this is illogical, indeed we think the goal is laudable, but rather that it presumes that constructs *can* be defined consistently and coherently across levels of analysis that span from genes, to molecules, to self-report, to lab tasks, without the recognition that at each level idiosyncrasies of methods give ample reason to be pessimistic. In other words, this can be understood as another manifestation or downstream consequence of the jingle fallacy. One cannot simply presume the same construct across methods, even if they have been similarly labelled. Serious research efforts must be undertaken to bridge constructs across methods *before* these constructs are used for prediction and explanation.

Measures first developed in small samples with relative impoverished psychometric models persist in fields as "gold-standard" measures due to researchers' familiarity rather than any evidence of quality. For example, a re-analysis of six large datasets on measures of executive function showed that the original factor structure reported by Miyake et al. (2000), studied in 137 college students, and cited over 13,000 times, did not outperform more standard and well-accepted models of cognitive function such as the Cattell-Horn-Cattell model (Jewsbury, Bowden, & Strauss, 2016). The NIH Toolbox measure of executive function included a single measure of discriminant validity (IQ), which was correlated at $r = 0.44–0.79$ across ages (Zelazo et al., 2013). "Grit" serves as another example; the original measure was so highly correlated with conscientiousness in the original paper ($r = 0.77$), that when corrected for unreliability it would approach 1.0 (Duckworth et al., 2007), not to mention serious critiques of the misapplication of factor analysis in that original manuscript (Credé, Tynan, & Harms, 2017). Without greater attention to the systematic description and careful, extensive measurement efforts, the field will continue to see the introduction, reification, and persistence of problematic measures.

In this way, we view the generalizability crisis described by Yarkoni to be a crisis of constructs. Behavioral scientists have a track record of subordinating external validity to internal validity, which is why we feel it important to highlight that business as usual is doing violence to both. The good news is the prescription is simple: the field should insist on, if not prize, a careful focus on methods and measures development, and deep construct validation. It is the bedrock of our science.

Financial support. Kevin King was funded by grants from NIDA (DA047247) and NIAAA (AA028832). Aidan Wright was funded by grants from NIAAA (AA026879).

Conflict of interest. None.

References

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*, 495–525. doi: 10.1037/a0019486

Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry. *American Journal of Psychiatry*, *171*(4), 395–397.

Jewsbury, P. A., Bowden, S. C., & Strauss, M. E. (2016). Integrating the switching, inhibition, and updating model of executive function with the Cattell–Horn–Carroll model. *Journal of Experimental Psychology: General*, *145*(2), 220.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH toolbox cognition battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, *78*(4), 16–33.

There is no generalizability crisis

Daniël Lakens[a], Duygu Uygun Tunç[b] and Mehmet Necip Tunç[c]

[a]Human-Technology Interaction Group, Eindhoven University of Technology, Atlas 9.402, 5600BM, Eindhoven, The Netherlands; [b]Philosophy Department, Middle East Technical University, Üniversiteler Mahallesi, Dumlupınar Bulvarı No:1, 06800 Çankaya/Ankara TURKEY and [c]Social Psychology Department, Tilburg University, Simon Building Room 405 Warandelaan 2 5037 AB, Tilburg.
D.Lakens@tue.nl
duygu.uygun@outlook.com
m.n.tunc@uvt.nl
https://sites.google.com/site/lakens2; https://uyguntunc.com/

doi:10.1017/S0140525X21000340, e25

Abstract

Falsificationist and confirmationist approaches provide two well-established ways of evaluating generalizability. Yarkoni rejects both and invents a third approach we call *neo-operationalism*. His proposal cannot work for the hypothetical concepts psychologists use, because the universe of operationalizations is impossible to define, and hypothetical concepts cannot be reduced to their operationalizations. We conclude that he is wrong in his generalizability-crisis diagnosis.

How generalizability claims should be justified has been a point of contention in psychology for decades. There are two well-established methodological perspectives on the issue. The falsificationist approach, a deductive strategy, consists of severely testing claims to discover the limits of their generalizability (Popper, 1959). The confirmationist approach, an inductive strategy, consists of accumulating single facts that collectively build partially confirmed generalizability claims (Carnap, 1936). Both approaches provide coherent and effective ways to evaluate generalizability claims in science. Yarkoni proposes a third approach built on the impossible ideal of *verifying* (i.e., conclusively confirming) generalizability claims through random-effect modeling. Unsurprisingly, he concludes that this approach is practically impossible to apply, because infinitely many factors exist that could moderate the generalizability of effects. Yarkoni's "crisis" narrative conflates his impossible approach to achieve a goal with the impossibility of achieving a goal. Generalizability claims are by definition based on extrapolation, and go beyond the data (Shadish, Cook, & Campbell, 2001). Generalizations are therefore always speculations on the basis of tentative assumptions that await falsification, or based on incrementally increasing beliefs through partial confirmations.

The falsificationist strategy is summarized by Mook (1983, p. 380): "We are not making generalizations, but testing them." Falsificationists test predictions of a theory, along with a ceteris paribus clause which posits that "nothing else is at work except factors that are totally random" (Meehl, 1990, p. 111). If the ceteris paribus clause holds, the claim is generalizable. Yarkoni is correct that ceteris paribus is often not literally true (cf. Meehl, 1990). Systematic non-trivial factors exist. However, all theories are necessarily simplifications: A map is never meant to be the territory (Bateson, 1972, p. 459). The challenge is to identify, from an infinite set of possible factors that falsify the theory's generalizability claim, which do so in a way that actually matters (Box, 1976). For example, although it is possible that temperature has a tiny impact on the Stroop effect, nobody considers it plausible that the effect will be meaningful enough to actually study it.

If experiments yield data that are too heterogeneous to be explained by a theory, either the theory or the ceteris paribus clause is falsified. If the latter option is chosen, a less general theory is proposed. Even when well-corroborated by the data, generalizability claims are only tentatively accepted. Lakatos (1978) reminds us that theories will always have unresolved problems, which is acceptable as long as our theories are "good enough" (Meehl, 1990, p. 115).

The confirmationist strategy is summarized by Carnap (1936, p. 425): "We cannot verify the law, but we can test it by testing its single instances. … If in the continued series of such testing experiments no negative instance is found but the number of positive instances increases then our confidence in the law will grow step by step." Within a confirmationist framework, researchers start by observing a single (often the most prototypical) instance of the investigated phenomenon. If subsequent observations enlarge the set of positive instances predicted by the theory, researchers increase their belief in its generalizability. Since verification is deemed impossible, confirmationists aim to specify the extent to which a generalizability claim is supported.

Yarkoni is not satisfied with either strategy and invents a third approach that we call *neo-operationalism*. Yarkoni's core argument is that generalizability claims need to be strictly data-driven, that is, based on random-effect modeling. He believes this is a feasible approach to "closely align" verbal and statistical hypotheses, which should lead to well-warranted generalizability claims.

His proposal cannot work for two reasons. First, we can only close the gap between concepts and their measures by stochastically sampling operationalizations from some underlying population if the meaning of the concept is identical to the population of its operationalizations. This is true for what MacCorquodale and Meehl (1948) call *abstractive* concepts, such as "color" in the Stroop task, which is identical to the colors in the visible spectrum. However, many concepts in psychology are *hypothetical* (e.g., "anger"), meaning they are semantically richer than and cannot be reduced to their operationalizations (e.g., anger is not just what anger measures measure). Thus, as long as psychologists want to theorize via hypothetical concepts, random-effect modeling cannot bridge the gap between verbal and statistical hypotheses, no matter how expansive the fitted model is (Green, 1992; Leahey, 1980).

Second, exhaustively defining a universe of operationalizations is impossible for hypothetical concepts (cf. Bear & Phillips commentary in this treatment): Such a "universe" would be too vast, theory-laden, and most probably time-dependent to be definable. Together, these points imply that statistical hypotheses will never be perfectly aligned with verbal hypotheses involving hypothetical concepts. Yarkoni's proposed solution could work if scientists limit themselves to abstractive concepts, but as Yarkoni still recommends the use of concepts such as "anger" or "charitable donation" in titles, which go well beyond any specific operationalizations, limiting psychological science to abstractive concepts seems too big a sacrifice.

Yarkoni's inductive neo-operationalism clearly does not sing from the same songbook as the deductive methodological falsificationist approach of tentatively accepting the ceteris paribus clause. But neo-operationalism is also in conflict with confirmationism, although both accounts are inductivist. For Yarkoni,

claims can only be generalized as far as they are aligned with the model that is fitted. By contrast, the confirmationist does not try to bridge the gap between verbal and statistical models. Partial confirmation is all one can get. To the extent that the generalizability claim is supported by novel data, the belief in it increases.

Yarkoni's recommendation to deal with the generalizability "crisis" is to clearly indicate that any extrapolation beyond the data is speculation (Sect. 6.3.1, para. 1). Rarely has there been a crisis solved more easily than by adding "going beyond the data" before a generalizability claim. Moreover, "going beyond the data" essentially means either tentative acceptance or partial confirmation. The diagnosis of a "crisis" is unwarranted when the two tried and tested approaches to justifying generalizability claims, the falsificationist and the confirmationist approach, already deliver what they promise. This leads us to conclude only one thing: There is no generalizability crisis.

Acknowledgment. Thanks to Leo Tiokhin for feedback on an earlier draft.

Funding. This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research, and by the European Union and the Turkish Scientific and Technological Research Council under the Horizon 2020 Marie Skłodowska-Curie Actions Cofund program Co-Circulation2.

Conflict of interest. None.

References

Bateson, G. (1972). *Steps to an ecology of mind*. Jason Aronson Inc.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791–799. doi: https://doi.org/10.1080/01621459.1976.10480949.

Carnap, R. (1936). Testability and meaning. *Philosophy of Science*, *3*, 419–471. doi: https://doi.org/10.1086/286432.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory & Psychology*, *2*(3), 291–320. doi: https://doi.org/10.1177/0959354392023003.

Lakatos, I. (1978). *The methodology of scientific research programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Leahey, T. H. (1980). The myth of operationism. *The Journal of Mind and Behavior*, *1*(2), 127–143. Retrieved from https://www.jstor.org/stable/43852818.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95–107. doi: https://doi.org/10.1037/h0056029.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141. doi: https://doi.org/10.1207/s15327965pli0102_1.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387. doi: https://doi.org/10.1037/0003-066X.38.4.379.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Publishing fast and slow: A path toward generalizability in psychology and AI

Andrew K. Lampinen, Stephanie C. Y. Chan
Adam Santoro and Felix Hill

DeepMind, London N1C 4DN, UK.
lampinen@google.com
scychan@google.com
adamsantoro@google.com
felixhill@google.com

doi:10.1017/S0140525X21000224, e26

Abstract

Artificial intelligence (AI) shares many generalizability challenges with psychology. But the fields publish differently. AI publishes fast, through rapid preprint sharing and conference publications. Psychology publishes more slowly, but creates integrative reviews and meta-analyses. We discuss the complementary advantages of each strategy, and suggest that incorporating both types of strategies could lead to more generalizable research in both fields.

The generalizability challenges outlined in the target article are not unique to psychology. Artificial intelligence (AI) – which also attempts to characterize and influence complex systems – is susceptible to many similar challenges. These include random effects of "subject" (random seeds), and unrecognized, unmeasured factors that affect conclusions (Bouthillier et al., 2021; Henderson et al., 2018; Weinberger, 2020). But the fields respond differently. Each field has different established practices on the publication or dissemination of research, and these different practices help to uniquely immunize the fields to some of these challenges. Could a publication strategy that incorporates elements from both fields be key to achieving generalizability?

In AI, publishing is rapid and multifaceted. Blog posts describe ideas before papers are written, and sharing pre-submission preprints on arXiv is standard practice. The vast majority of novel empirical findings, whether incremental or paradigm-altering, either remain as preprints or are rapidly published in the peer-reviewed proceedings of annual conferences, rather than journals.

Publishing fast accelerates progress in AI. It allows authors to get rapid, broad feedback, and encourages early discovery of the settings where ideas do or do not generalize. Faster publishing is also more inclusive – preliminary knowledge is shared with the entire community, rather than only those who happen to know the author, or who can afford to subscribe to the right journals or attend the right conferences.

In psychology, publishing is slower. Articles are longer, typically summarizing the results of a series of closely-related experiments. In an even slower process, articles are aggregated into larger reviews and meta-analyses.

Publishing slowly allows psychology to carefully explore phenomena, and to integrate the results of many studies. While the writing in individual articles may elide important factors of variation, as cautioned by the target article, psychology studies include more carefully controlled manipulation of some factors than studies in AI. Meta-analyses and reviews attempt to fill the gaps, outlining the limits of a phenomenon and integrating related works, as do journals (like this one) that explicitly encourage debate. Psychology values broader analyses and summaries as an important part of scientific research. Summaries also increase inclusivity, by making the state of knowledge readily available to those who are not directly immersed in the literature or community.

However, publishing fast and slow need not be mutually exclusive. Their benefits are complementary, and each field could learn from the other.

Psychology should incorporate faster publishing, including early preprints, dataset sharing, and conference publications. If researchers shared more preliminary and negative results (as preprints and conference papers), the field could more rapidly learn which factors of variation might be important moderators of an effect. By contrast, relying on journal publications delays the dissemination of research, and increases the bias toward positive results. Fast publishing reduces pressure to present fully developed, distinct stories, instead favoring incremental developments and collaboration across the community. This relates to dataset sharing, which has helped AI to progress, as noted in the target article (section 6.3.7). Collecting a new dataset for each paper slows the development and sharing of research. Thus, we agree that shared datasets – as well as experiment code and materials – help improve the generalizability of research.

It may seem counterintuitive to suggest that psychology should accelerate publishing, given recent arguments for more careful deliberation, even including replication and meta-analysis within papers (McShane & Böckenholt, 2017). What AI shows, however, is that ideas are more thoroughly explored by engaging the broader research community. The ultimate construction of an overarching theory should aggregate across many papers, produced by many unique groups, each with their own biases, apparatuses, and experimental techniques. Fast publishing thus seeds slow publishing; rapidly producing varied studies around a conceptual theme provides the basis for more generalizable summaries. High-level hypotheses and arguments that are not yet sufficiently supported can be shared in blog posts. Thus, individual experimental papers (especially preliminary preprints) can state more conservative, descriptive conclusions, as the target article suggests, but broader speculation and extrapolation can nevertheless be shared.

AI should incorporate slower publishing, including integrative reviews and meta-analyses. Fast publishing in AI often leads to communal knowledge about which techniques are beneficial in which settings. But this knowledge is rarely integrated or made explicit. AI would benefit from more reviews and meta-analyses that quantify variance components across many experiments (target article section 6.3.4–5). There is increasing evidence that unmeasured factors affect conclusions in AI, for example, environmental realism and embodiment (Hill et al., 2020), or reward scales and random seeds (Henderson et al., 2018). These factors can be partially addressed by more careful experimentation and statistics (Henderson et al., 2018; Weinberger, 2020), especially accounting for random effects. However, no single paper can explore every factor, so aggregating research is critical to achieving generalizable understanding.

While incorporating both fast and slow publishing would help, this strategy comes with challenges. Implementing it would require altering incentive structures. Psychology would need to recognize the value of imperfect preprints as research contributions. AI would need to value summarizing articles, even if their primary contribution is to clearly articulate and evaluate common knowledge within the field, rather than proposing something new. Finally, the public and press would need to avoid overinterpreting preliminary results.

There are also research challenges. While shared datasets can accelerate publishing, a "dataset-as-fixed-effect" fallacy can reduce generalizability. For example, common techniques that improve ImageNet performance are detrimental on other datasets because they bias models to rely on texture (Hermann, Chen, & Kornblith, 2020). There must be sufficient dataset diversity to ensure that the entire community does not overfit (Grootswagers & Robinson, 2021). But it is easier to identify and correct these issues by exploring, sharing results quickly, and integrating knowledge across many studies.

In summary, generalizability in both psychology and AI would be improved if both fields embraced two "systems" of publication: one rapid and reactive, and the other slower and more deliberate. By rapidly exploring many variations on an idea, and then integrating the results through broad meta-analyses and reviews, both fields could more efficiently arrive at generalizable insights about their domains of inquiry.

Acknowledgments. We thank Julie Cachia and Yochai Shavit for comments on this manuscript.

Financial support. The authors are funded by DeepMind.

Conflict of interest. None.

References

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., … Vincent, P. (2021). Accounting for variance in machine learning benchmarks. *arXiv preprint* arxiv:2103.03098.

Grootswagers, T., & Robinson, A. K. (2021). Overfitting the literature to one set of stimuli and data. *arXiv preprint* arXiv:2102.09729.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 3207–3214). February 2–7, 2108. New Orleans, Louisiana.

Hermann, K. L., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/pdf?id=SklGryBtwr.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research*, *43*(6), 1048–1063.

Weinberger, K. (2020). On the importance of deconstruction. Presentation, NeurIPS 2020 Retrospective Workshop.

Is formalism the key to resolving the generalizability crisis? An experimental economics perspective

Zacharias Maniadis

Economics Department and Center for Economic Policy Research, University of Cyprus, CY-1678 Nicosia, Cyprus and Department of Economics, School of Social Sciences, School of University of Southampton Highfield Campus, Southampton SO17 1BJ, UK.
z.maniadis@soton.ac.uk; https://sites.google.com/site/zachariasmaniadiswebpage/

doi:10.1017/S0140525X21000273, e27

Abstract

I draw lessons from experimental economics. I argue that the lack of mathematical formalism cannot be usefully thought as the cause of the underappreciation of contextual and generalizability considerations. Instead, this lack is problematic because it hinders a clear relationship between theory and quantitative predictions. I also advocate a pragmatic policy-focused approach as a partial remedy to the generalizability problem.

Formalism is not associated with a strong focus on context

Yarkoni correctly points out that the practice of using verbal models tested by statistical tools can be problematic. It then elaborates, focusing almost exclusively on the issue of context-dependence of human behavior and the overall complexity of the subject matter of social science, which renders generalizability of research findings difficult. I wish to contribute to the discussion from the perspective of a different behavioral discipline, experimental economics, which uses the mathematical language in its theory more frequently than psychology. I draw lessons from that discipline to show that the elaborated generalizability issues, while relevant, are not related directly to the lack of formalism.

In experimental economics, theories are predominantly mathematical, rather than verbal. Economics employs a set of principles, based on which, deductive models are constructed. These include preferences, beliefs, optimization, and equilibrium. Models and their associated properties are pitted against each other using data, while the formal rigor facilitates clear connection among models, underlying principles, and empirical methods. Prominent scholars have long regarded the assessment of competing models, not external validity, as the main focus of experiments (Plott, 1982; Smith, 1976). Schram (2005) argued that "external validity has received much more attention in psychology than in economics. To a large extent, psychological research is inductive and based on observed empirical regularities."

Camerer (2011) clearly explains why experimental economics has traditionally had a weaker concern for generalizability to real-life settings: "all empirical methods are trying to accumulate regularity about how behavior is generally influenced by individual characteristics, incentives, endowments, rules, norms, and other factors. A typical experiment therefore has no specific target for 'external validity'…." According to this view – called the "scientific" view – a theory-testing experiment helps choose between different theories and connects to our current understanding of the world.

Partly because of this specific methodological tradition, the issues that Yarkoni develops in the main text have not received major attention in experimental economics. As Loewenstein (1999) and Levitt and List (2007) argue, external validity or sampling concerns have not been given more focus relative to psychology – but see Exadaktylos, Espín, and Branas-Garza (2013) – and contextual variables are not regularly incorporated in models as Yarkoni envisions. Duflo (2017) argues: "details that we as economists might consider relatively uninteresting are in fact extraordinarily important in determining the final impact of a policy or a regulation, while some of the theoretical issues we worry about most may not be that relevant."

A literature comparison indicates that experimental economists do not introduce and systematically vary contextual factors more frequently than psychologists (especially within a given study, as Yarkoni advocates). Because of their interest in general principles, economists focus more on the importance of homogenizing important types of stimuli and removing context (Hertwig & Ortmann, 2001). However, Levitt and List (2007) argue that cross-situational consistency of behavior is lacking, which requires theories and methodologies to be addressed (e.g., see Galizzi & Navarro-Martinez, 2019). Coinciding with a possible reproductivity crisis in science (see Ioannidis, 2005; Maniadis, Tufano, & List, 2014), theoretical interest in generalizability has recently increased (Kessler & Vesterlund, 2015; List, 2020; Zizzo, 2013).

To summarize the point: for experimental economics, it is not the case that the use of mathematical theories for decades was accompanied by a focus on the importance of heterogeneity of stimuli and other contextual factors. Instead, formal theory-testing is considered a domain where generalizability concerns should apply less. The problem of context-dependence in psychology may deserve to be addressed by careful statistical models and advanced experimental designs. However, the verbal representation of theories does not seem to be the culprit.

Advantages of formal theory

I argue that the lack of formal theories in psychology is more problematic for another reason: it hampers clear theoretical predictions. In economics, formalism facilitates a relatively tight logical connection between theory and predictions. Accordingly, statistical research hypotheses follow theory naturally. Hence, it is more difficult to account – using ad hoc arguments – for experimental evidence inconsistent with a given theory. Muthukrishna and Henrich (2019) and Ortmann (2020) also advocate mathematical formalism to help us understand what a theory predicts and what it does not.

Contrary to the main connection made in Yarkoni, a formal framework grounded on a set of overarching principles may facilitate knowledge accumulation not by allowing an arbitrary number of moderators to be considered, but by restricting the set of questions that are considered reasonable. This aspect of theory in experimental economics is now attracting some attention in psychology (Muthukrishna & Henrich, 2019). However, one needs to be cautious: while formalism makes excessive ad hoc theorizing more difficult, it does not rule it out.

Experimental economics seems to fare better in terms of replicability (Camerer et al., 2016), and rigorous theory plays a role in this. However, this rigor mediates replicability primarily via some of the secondary channels mentioned in Yarkoni: making riskier predictions and explicitly comparing competing theories. Predictions in economics tend to be much more quantitative and often estimation (rather than statistical hypothesis-testing alone) is the objective.

A pragmatic approach

If the target is applicability to specific domains rather than theory-testing, another approach could be used. Randomized controlled trials in development and public economics examine the performance of interventions in natural environments. This methodological approach has been compared to that of plumbers, dentists, or engineers (Duflo, 2017; Roth, 2002, 2018), and may be useful as a partial remedy to a possible "generalizability crisis." Variability-enhancing designs that examine a high number of psychological factors may not always be pragmatic or feasible. Instead, in many cases of interest, one could focus on specific policy domains and try to emulate them. A promising approach is assessing systematically whether the effect size of a given intervention is robust to the intervention being scaled-up as a full policy (Al-Ubaydli, List, & Suskind, 2017). Acknowledging the importance of scalability in concrete policy domains could be a less ambitious – but potentially useful – approach for addressing a potential generalizability crisis.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2017). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, *107*(5), 282–286.

Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. Available at SSRN 1977749.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351* (6280), 1433–1436.

Duflo, E. (2017). Richard T. Ely lecture: The economist as plumber. *American Economic Review*, *107*(5), 1–26.

Exadaktylos, F., Espín, A. M., & Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports*, *3*(1), 1–6.

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: A systematic lab-field study. *Management Science*, *65*(3), 976–1002.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists?. *Behavioral and Brain Sciences*, *24*(3), 383–403.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kessler, J., & Vesterlund, L. (2015). *The external validity of laboratory experiments: The misleading emphasis on quantitative effects* (Vol. *18*, pp. 392–405). Oxford, UK: Oxford University Press.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives*, *21*(2), 153–174.

List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial (No. w27535). National Bureau of Economic Research.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, *109*(453), F25–F34.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, *104*(1), 277–290.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Ortmann, A. (2020). On the foundations of behavioural and experimental economics. Available at SSRN.

Plott, C. R. (1982). Industrial organization theory and experimental economics. *Journal of Economic Literature*, *20*(4), 1485–1527.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, *70*(4), 1341–1378.

Roth, A. E. (2018). Marketplaces, markets, and market design. *American Economic Review*, *108*(7), 1609–1658.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225–237.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274–279.

Zizzo, D. J. (2013). Claims and confounds in economic experiments. *Journal of Economic Behavior & Organization*, *93*, 186–195.

The "'Crisis' Crisis" in psychology

John D. Medaglia[a,b,c] and Kiante A. Fernandez[d]

[a]Department of Psychological and Brain Sciences, Drexel University, Philadelphia, PA, 19104, USA; [b]Department of Neurology, Drexel University, Philadelphia, PA, 19104, USA; [c]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19119, USA and [d]Department of Psychology, Ohio State University, Columbus, OH 43210, USA.
jdm582@drexel.edu; kaf395@drexel.edu; www.cognew.com

doi:10.1017/S0140525X21000364, e28

Abstract

The recent trend to label dilemmas in psychology as "crises" is insidious. The "'Crisis' Crisis" in psychology can distract us from actionable practices. As a case in point, "The Generalizability Crisis" offers the valuable central thesis that verbal-quantitative gaps imperil psychological science. Focusing on the key issues rather than crisis narratives can lead to progress in our discourse and research.

"The Generalizability Crisis" offers lucid insights into the problems that can occur when researchers inappropriately use statistical models to test hypotheses and generalize their findings. However, labeling trends in psychological science as "crises" is a new form of insidious professional communication. Authors can convey a sense of immediacy, severity, and unity of causes and effects that are sensationalizing at best and counterproductive at worst. Collective action could mitigate the trends that ultimately lead authors to write "crisis" articles. Psychologists have possessed the intellectual and quantitative tools to prevent these dilemmas in psychological science for many years. We should expect that people will often fail to apply these tools (Lilienfeld, 2017). Furthermore, we should not expect powerful new analyses to automatically solve all problems (Kell & Oliver, 2004; Shneiderman, 2016). With the intent to halt an emerging rhetorical practice, we concisely call attention to the "'Crisis' Crisis" in psychology, its possible causes, and how to obviate it.

Crises are turning points: unstable or critical moments in which a decisive change is impending (Merriam-Webster, n.d.), often associated with difficulty, danger, and suffering (Dictionary, n.d.). It is not clear that this term best describes the concerns offered by Yarkoni, nor the preceding and ongoing *replication crisis* (Maxwell, Lau, & Howard, 2015) or *reproducibility crisis* (Baker, 2016; Open Science Collaboration, 2015). Recently, *crisis narratives* have been uncritically endorsed by most authors writing about the topic, whereas the trends in question might be better characterized as epochal change (Fanelli, 2018). With the increased availability of meta-research practices (Lakens, Hilgard, & Staaks, 2016; Soderberg et al., 2020), transparent and open data sharing and preregistration (Nosek, Ebersole, DeHaven, & Mellor, 2018), and instant and accessible communication technologies, enhanced visibility and discussion of undesirable practices could be a harbinger of positive change, not a crisis to be averted (Nosek et al., 2021). By espousing crisis narratives, we should be mindful of the risk of contributing to bandwagoneering negativity, cynicism, indifference, and antiscientific sentiments (Fanelli, 2018). Considered unironically, the "'Crisis' Crisis" in psychology is no crisis at all if we identify its causes and obviate further alarm.

An unfortunate victim of the "'Crisis' Crisis" is one of Yarkoni's excellent central theses. We could not agree more that the gulf between verbal statements and inferential statistics (or any quantitative concept) can impede progress in psychology. On that basis, we fully agree with Yarkoni's suspicion that the verbal-quantitative divide in psychology is one of its fundamental challenges. Unfortunately, this long overdue critique is wrapped within the narrative-reinforcing guise of a crisis, which takes some of the emphasis off of the key argument. Nontrivially, the verbal-quantitative divide is philosophically and temporally antecedent to any specific concerns about generalizability. The verbal-quantitative divide is not unique to issues that cause or contribute to problems that arise when psychologists aim to generalize their findings. Nor is it isolated to issues inherent in applying linear mixed models to psychological and behavioral data, or other cases that Yarkoni considers. This basic point is so significant that we were motivated to write this response article to urge authors to eschew "crisis" overtones. In this case, we should instead focus on a more fundamental issue that affects much of psychological science: too often, what we say we study does not match what we do quantitatively. That is a big problem.

How did we arrive at this crisis of all crises in psychology? Authors often respond to trends that have a basis in facts, but the tendency to call them "crises" is worth a moment's reflection.

If crisis narratives are self-reinforcing with relatively few published or conversational counterpoints, they may be trivially self-sustaining. As others have noted, the concern that research quality is declining is neither new nor universally justified (Fanelli, 2018), and calling attention to misapplied concepts and methods is part of the routine business of science (Gelman & Loken, 2016). Perhaps psychologists are alarmed when they notice patterns and trends that they did not perceive before due to our increased focus on meta-research. Others could be concerned that psychological science is perverse incentives or bad actors all the way down (Lilienfeld, 2017). In either case, scientists have created the tools to identify, evaluate, and communicate about these trends, call attention to them, and propose solutions. Perhaps the perceived value of calling for a virtuous change in academic psychology has increased such that some authors spend time writing about it (Whitaker & Guest, 2020). Labeling calls to change a "crisis" could receive more press, views, and citations, which can be potent reinforcers for authors and editors (Dworkin et al., 2020; King, Bergstrom, Correll, Jacquet, & West, 2017; Moed et al., 2012; Ruscio, 2016).

Systemic problems call for collective and individual actions. To identify and absolve "crisis"-driving issues, we point readers to copious writing about incentives and other structural issues including salary composition (Bourne, 2018), research funding dynamics (Lilienfeld, 2017; Wahls, 2018), citation practices (Stephan, 2012), and individual researcher choices (Chambers, 2017). Special attention to reinforcing best standard practices for established and new quantitative practices is essential (Amrhein, Trafimow, & Greenland, 2019; Loken & Gelman, 2017; Shrout & Rodgers, 2018; Trafimow, 2018). Many issues would be resolved by adhering to practices we teach to students (Chopik, Bremner, Defever, & Keller, 2018). Others would be resolved by consulting quantitative domain experts who can help bridge verbal-quantitative gaps. In agreement with Yarkoni, we encourage readers to ask if a concept is quantifiable, and if so, find a procedure to sufficiently test the idea. We should always ask what we aim to study, how to measure it, and what we can and cannot conclude from our experimental and quantitative methods. If we must consider a qualitative method, we should still justify it and consider what quantification could benefit us. It could be that psychologists generate many verbalizable concepts that are invalid and worth no further pursuit. But we should remember that most sciences have been in this predicament at some point in time.

Financial support. JDM acknowledges support from NIH grants DP5-OD-021352-01, R01-DC-16800-01A1, R01-DC-014960-01A1, R01-AG-059763, and Department of the Army grant PRMRP 12902164.

Conflict of interest. None.

References

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(S1), 262–270.

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353–366.

Bourne, H. R. (2018). Opinion: Expansion fever and soft money plague the biomedical research enterprise. *Proceedings of the National Academy of Sciences*, *115*(35), 8647–8651.

Chambers, C. (2017). *The seven deadly sins of psychology*. Princeton University Press.

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*(2), 158–163.

Dictionary, C. (n.d.). Crisis. In dictionary.cambridge.org dictionary. Retrieved February 13, 2020 from https://dictionary.cambridge.org/us/dictionary/english/crisis.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, *23*(8), 918–926.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*(11), 2628–2631.

Gelman, A., & Loken, E. (2016). The statistical crisis in science. In M. Pitici (Ed.), *The best writing on mathematics* (pp. 305–318). Princeton University Press.

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, *26*(1), 99–105.

King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, *3*, 2378023117738903.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 1–10.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

Merriam-Webster. (n.d.). Crisis. In Merriam-webster.com dictionary. Retrieved February 13, 2020 from https://www.merriam-webster.com/dictionary/crisis.

Moed, H. F., Colledge, L., Reedijk, J., Moya-Anegon, F., Guerrero-Bote, V., Plume, A., & Amin, M. (2012). Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics*, *92*(2), 367–376.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., … Vazire, S. (2021). Replicability, robustness, and reproducibility in psychological science. Retrieved from https://psyarxiv.com/ksfvq/.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Ruscio, J. (2016). Taking advantage of citation measures of scholarly impact: Hip hip h index!. *Perspectives on Psychological Science*, *11*(6), 905–908.

Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, *113*(48), 13538–13540.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.

Soderberg, C. K., Errington, T., Schiavone, S. R., Bottesini, J. G., Thorn, F. S., Vazire, S., … Nosek, B. A. (2020). Research quality of registered reports compared to the traditional publishing model. Retrieved from https://osf.io/aj4zr/.

Stephan, P. (2012). Perverse incentives. *Nature*, *484*(7392), 29–31.

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology*, *31*(8), 1188–1214.

Wahls, W. P. (2018). Point of view: The NIH must reduce disparities in funding to maximize its return on investments from taxpayers. *Elife*, *7*, e34965.

Whitaker, K., & Guest, O. (2020). #bropenscience is broken science: Kirstie Whitaker and Olivia guest ask how open "open science" really is. *The Psychologist*, *33*, 34–37.

Psychologists should learn structural specification and experimental econometrics

Don Ross[a,b,c]

[a]School of Society, Politics, and Ethics, University College Cork, Cork, T12 AW89, Ireland; [b]School of Economics, University of Cape Town, Rondebosch 7701, South Africa and [c]Center for Economic Analysis of Risk, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303, USA.
don.ross931@gmail.com; http://uct.academia.edu/DonRoss

doi:10.1017/S0140525X21000108, e29

Abstract

The most plausible of Yarkoni's paths to recovery for psychology is the least radical one: psychologists need truly quantitative methods that exploit the informational power of variance and heterogeneity in multiple variables. If they drop ambitions to explain entire behaviors, they could find a box full of design and econometric tools in the parts of experimental economics that don't ape psychology.

The methodological tradition of experimental design and analysis in psychology has generated epistemological pathology that undermines the discipline. The crisis is more serious because it is not recent and acute but old, deep, and chronic. Yarkoni's proposed responses invite very different metrics of assessment. Abandoning standard experimental psychology would obviously pitch out babies with bathwater, but trying to estimate the babies/bathwater ratio would be daunting. Resorting entirely to qualitative description and reflection would be only a slower and less transparent path to disciplinary suicide. If, as Yarkoni argues, most of the qualitative relationships that psychologists treat as hypotheses *obviously* obtain *some of the time*, then it is hard to see why we would want to maintain a whole academic discipline merely to pronounce these truisms. We already have other people better trained to unearth surprising implications of apparently protean psychological truths, namely philosophers. I therefore prefer to focus on Yarkoni's proposed path to better quantitative practice. I suggest that (some) economists offer a useful model of practice here.

I will start with a philosophical question: What exactly is experimental psychology supposed to be *for*? A possible if imprecise answer, which none of Yarkoni's rhetoric seems to contest, is "explaining and predicting behavior that results from biological information processing." But we might object that almost no behavior by such an incorrigibly social and culturally embedded species as humans results *only* from biological information processing. This applies to even the simplest behaviors. Suppose you wanted to model the production of conversation-supporting hand gestures in a group of speakers. Some gestures will have been developed by individuals as solutions to communicative aims they idiosyncratically find difficult. More will have been copied from specific role models. Most will simply have been inherited from childhood cultural learning samples. These three kinds of data-generating processes (DGPs) involve varying combinations of psychological, cultural-environmental, and economic causal pathways. Now, also "obviously," biological information processing plays an essential *role* in all three: social influences have to influence motor control systems in speakers' brains. If we seek a general theory of hand gesturing, then the contribution we ideally want from the psychologist here is to *partial out* this component of the overall behavior-generating process. This could mean (qualitatively) locating it in the flow-chart diagram of a putative mechanism, or (quantitatively) assigning a parameterized weight to a coefficient associated with it. By illustrative contrast, the economist's job is to figure out, for example, how relatively entrenched different gestures are in response to shifting incentives around trade-offs between signal precision and cross-audience effectiveness.

I think that the triumph of connectionist over classical artificial intelligence showed us the naïveté of the boxological approach. This is why Yarkoni's third response to the methodological crisis is where the action is. But then clearly one shouldn't try to implement it by seeing whether one can statistically reject a hypothesis that includes the assumption that everything in each DGP except the encapsulated dynamics in speakers' nervous systems, and indexicals for their specific histories, is a fixed effect in a linear model. Of course one can generate data to reject any hypothesis in this class, because they're all certainly false. Piling up such rejections takes one not a jot closer to a general model, both because the process of elimination is endless but also because the entire search takes place in the wrong solution space.

Economists are also in the business of trying to partial out one element in the behavior production function, marginal incentive changes. But because they know that incentives operate through multiple channels, including channels where information relevant to successful goal achievement are typically hidden from the subject (she copies successful bond investors, she doesn't simulate them), they are less likely to have strong priors on what might constitute a "confound." Indeed, those experimental economists who have not, disastrously, aped psychological methodology (as have, alas, the behavioral economists who most beguile non-economist audiences) don't tend to use the language of "confounds" at all. They assume that everything on the right-hand side of a structural model specification that might vary needs its own treatment group. Instead of trying to wall extra-economic causal factors out of the lab, they expand (in effect) the boundaries of the lab under theoretical guidance. This is expensive. Fortunately, the prevailing funding ecosystem in economics has evolved a norm that good experiments usually need generous budgets.

None of this would help much if econometric estimation techniques didn't co-evolve with the complexity of the structural models that are used as identification templates in lineages of experiments. But, at least since the coming of lots of cheap data-processing power, such co-evolution has been supported by the division of labor and resources in the discipline. Yarkoni's third path would have psychologists embracing the informational power of variance and heterogeneity, both in the kinds of model specifications they build and in their experimental designs. So, I suggest, they should study Bayesian experimental econometrics (Andersen, Harrison, Lau, & Rutström, 2010; Kruschke & Liddell, 2018; Lee & Wagenmakers, 2013). If they designed experiments so as to make full use of the resulting expansion of inferential power, individual experiments would become significantly more expensive, even in the absence of economists' special reason for needing to generously pay subjects. But it seems clear that Yarkoni agrees that a world with many fewer but much better psychological experiments would be an improved world.

A critic might reject my "partialling out" job description for psychology, and insist that psychologists aim to explain entire behavioral complexes rather than the aspects of behavioral causation that are psychological. Such hubris would greatly raise the stakes of Yarkoni's challenge: a claim that all of behavioral science should follow dead-end methodology would deserve the fiercest resistance. Economists have frequently had to unlearn disciplinary imperialism the hard way, but most now acknowledge that although people respond to economic incentives, they aren't ruled by them. Psychologists should be motivated to disciplinary modesty by the same kind of shock that eventually reformed standard practice in economics, failure to actually accumulate knowledge.

Financial support. This research received no specific grant from any funding agency.

Conflict of interest. None.

References

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2010). Behavioral econometrics for psychologists. *Journal of Economic Psychology, 31*, 553–576.

Kruschke, J., & Liddell, T. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review, 25*, 178–206.

Lee, M., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

The crisis from above: Gatekeepers need better standards

Sarah R. Schiavone[a], Julia G. Bottesini[a] and Simine Vazire[b]

[a]Department of Psychology, University of California, Davis, CA 95616, USA and [b]Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, VIC, 3010, Australia.
sschiavone@ucdavis.edu, https://sschiavone.com/
jbottesini@ucdavis.edu, https://juliabottesini.wordpress.com/
simine.vazire@unimelb.edu.au, http://simine.com

doi:10.1017/S0140525X21000546, e30

Abstract

Improvements to the validity of psychological science depend upon more than the actions of individual researchers. Editors, journals, and publishers wield considerable power in shaping the incentives that have ushered in the generalizability crisis. These gatekeepers must raise their standards to ensure authors' claims are supported by evidence. Unless gatekeepers change, changes made by individual scientists will not be sustainable.

Yarkoni's sobering description of the state of psychological science should make us all uncomfortable. While Yarkoni acknowledges that the ongoing problems undermining the validity of research are structural and reflect the norms of the field, his proposed courses of action focus nearly exclusively on individual researchers, and how they can improve the quality of their own research. However, there are pivotal players in shaping psychological science whose influence should likewise be addressed: the gatekeepers (i.e., journals, editors, publishers, and society board members). Not only do they bear much of the responsibility for the current state of affairs, but much of the change that Yarkoni (and we) desire would follow swiftly if a small group of gatekeepers decided to make it a priority. Moreover, unless the gatekeepers change, changes made by individual scientists will not be sustainable. Thus, targeting gatekeepers when calling for reform is not only more just, but a more practical avenue for achieving long-term change.

Relying on individual researchers' initiative to decide to take the harder road is not enough. Some motivated researchers will indeed heed the call, take pains to improve the generalizability of their findings, and rein in their conclusions to better correspond with their evidence. Ideally, these individuals would be rewarded for doing so. In reality, they may not be – and some journals, editors, funders, and committee members may find such calibrated claims to be underwhelming. Maybe the field improves slightly, but most likely incentive structures remain unchanged and consequently, many of those motivated researchers on the job market may find themselves passed over in favor of candidates who followed the traditional road of sweeping generalizations.

Gatekeepers such as journal editors, society board members, and publishers, on the other hand, are in a safe position to disrupt the status quo and change the standards for what counts as excellent research. These decision-makers set policies that determine what factors are weighed in journal acceptance – the currency by which researchers are evaluated. If the leading journals decided to raise their standards on a particular dimension, such as generalizability, researchers would be motivated to meet this new standard. Journals would have nothing to lose, unless they fear that their reputations depend on maintaining low standards.

Yarkoni suggests that we do not judge anyone too harshly for choosing the road of business as usual given the norms that have long been embraced in psychology. We acknowledge the unfairness in holding individual researchers to higher standards than those applied in the field at large. However, "ignore[ing] the bad news" should not be a viable option for those who wish to publish in our top journals – and certainly not for those who run them. Rather, journals (and the editors leading them) should be judged harshly if they repeatedly demonstrate a lack of concern about the generalizability of the findings they publish, and continue to reward novelty over rigor. If we cannot expect (or even demand) this from the very gatekeepers who shape incentives in the field, presumably valued for their ability to identify high quality work, what can we expect from them?

We agree wholeheartedly with Yarkoni that "Researchers must be willing to look critically at previous studies and flatly reject – on logical and statistical, rather than empirical, grounds – assertions that were never supported by the data in the first place, even under the most charitable methodological assumptions" (sect. 5, para. 3). We simply believe that the onus is primarily on those who control the biggest rewards – acceptance into the field's top journals – to put these practices into action. If we are going to ask this of researchers, we should not hesitate to expect the same (and arguably more) of editors, such as not letting such studies through peer review in the first place.

There are several paths journals can take moving forward. First, they could "raise the bar" and require authors to improve the quality of their research to support the sorts of claims that the field has traditionally enjoyed making. Many journals are already quite selective, but do not place sufficient weight on the validity of the methods and inferences when making editorial decisions. This could be addressed by paying methodologists and statisticians to serve as expert reviewers and encouraging registered reports (to provide feedback and identify problems prior to data collection, when they can still be addressed).

Second, journals could require claims to be limited to only what the research supports and accept that discussion sections will be far less spectacular. Statements on limitations and constraints on generality (Simons, Shoda, & Lindsay, 2017) should not be treated as confessionals to be buried within discussion sections and otherwise never spoken of again. Rather, claims made throughout an article should be expected to align with these statements. Press releases should similarly be written in ways that are

compatible with the strength of the evidence in the paper, even if that means they will garner less attention from the press, policymakers, and the public. Although some may mourn the decline in attention and influence, the credibility that this would afford the field would be well worth the cost. If journals want to allow authors space to speculate beyond what they have evidence for, these speculations should be relegated to a clearly-marked section, and should not find their way into abstracts, conclusions, and press releases.

We can not force journals to take any of these steps. Although we can urge them to do so, they may choose to simply carry on with business as usual – enforcing haphazard standards, rewarding novelty, and maintaining the "kind of collective self-deception" Yarkoni described. In that case, however, they should admit that what they are doing is not science (Campbell, 1984). Moreover, editors who continue to give stamps of approval to articles riddled with unsubstantiated claims should strongly consider Yarkoni's first proposed course of action: doing something else.

Acknowledgements:. None.

Financial support. This work was supported by a NSF Graduate Research Fellowship (#1247392) to Sarah R. Schiavone.

Conflict of interest. None.

References

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Connor, D. G. Altman, & C. Jackson (Eds.), *Evaluation studies review annual* (Vol. 9, pp. 26–48). Newbury Park, CA: Sage.

Simons, D. J., Shoda, Y., & Lindsay, S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. doi:10.1177/1745691617708630

Causal complexity demands community coordination

Beau Sievers and Evan DeFilippis

Department of Psychology, Harvard University, Cambridge, MA 02138, USA.
beau@beausievers.com; defilippis@g.harvard.edu

doi:10.1017/S0140525X21000418, e31

Abstract

Yarkoni's argument risks skepticism about the very possibility of social science: If social phenomena are too causally complex, normal scientific methods could not possibly untangle them. We argue that the problem of causal complexity is best approached at the level of scientific communities and institutions, not the modeling practices of individual scientists.

The argument advanced by Yarkoni (this issue) is, at its core, the problem of provisos, a skeptical dilemma in the philosophy of science posed by Hempel (1988) and Lange (1993): Claims that require *ceteris paribus* assumptions (i.e., that hold only "all else being equal") either overgeneralize, in which case they are false, or they depend on an unbounded set of provisos, in which case they are vacuous. What distinguishes Yarkoni's account from prior formulations is his simultaneous embrace of both horns of the dilemma: Scientists must fully align their verbal and statistical claims, and either abandon generalization or use very large models that favor prediction over explanation.

Building models large enough to match verbal claims might be quite difficult, and the target article says very little on how to go about it. Disturbingly, we are warned that model growth could be unbounded – we should be prepared to add factors "ad infinitum," with the understanding that "reality is not under any obligation to only manifest sparse causal relationships that researchers find intuitive." Without qualification, this opens the door to two more skeptical challenges: underdetermination and the problem of unconceived alternatives. If the true causal structure of reality is complex and unbounded, then any empirical finding could be caused by more than one independent mechanism, and all scientific theory would be underdetermined by the available evidence (Quine, 1951). Worse, if the nature of causality is inaccessible to human intuition, then even if we cobble together an instrumentally useful theoretical framework, we should expect it to be undermined by alternatives that we could not possibly conceive of (Stanford, 2006).

Following these skeptical implications, it seems to us that the true antagonist of Yarkoni's story is not the unfortunate tendency of social scientists to make sloppy verbal generalizations, but rather the complex causal structure of reality itself. Better alignment between verbal and statistical claims may make the problems of causal complexity easier to see, but it will not do much to solve them. Where Yarkoni asks why social scientists are so slapdash in their verbal treatment of statistics, we would ask instead: Are the institutions of social science competent to manage the challenges posed by the causal complexity of social phenomena?

One reason for optimism is that social scientists do occasionally come to generalizable conclusions without large statistical models or full alignment of verbal and statistical claims. For example, Knetsch and Sinden (1984) speculated that the endowment effect would generalize broadly, despite using a model that did not permit generalization beyond students at the University of New England trading lottery tickets for $3 in the year 1984. Later research iteratively modified this generalization scope, showing that the endowment effect extends across goods (e.g., Rowe, D'Arge, & Brookshire, 1980; van Dijk & van Knippenberg, 1998), age groups (Harbaugh, Krause, & Vesterlund, 2001), cultures (Maddux et al., 2010), and species (Lakshminaryanan, Keith Chen, & Santos, 2008), but is limited by market features (Kahneman, Knetsch, & Thaler, 1990), learning (Apicella, Azevedo, Christakis, & Fowler, 2014; Coursey, Hovis, & Schulze, 1987), and expertise (List, 2003). At the risk of overgeneralizing from a single example, this shows that the familiar, messy process of iterative science is in some cases capable of untangling the causal structure of social phenomena. Importantly, this process depends on scientists pragmatically understanding over- or under-generalization as opportunities to contribute. Nuanced research cannot get off the ground if there is nothing to nuance.

But how should scientists identify unmeasured factors? The target article recommends "careful, critical thinking," which is somewhat vague. In practice, this critical thinking does not take place inside the head of a single scientist, but through dialog within a broad community of scholars with diverse backgrounds and areas of expertise. Accordingly, the diversity of the scientific community

may be a better target for intervention than statistical modeling practice. Achieving what Harding (1992) calls "strong objectivity" will require creating variation not only in experimental factors, but also in the social standpoints of the people who select those factors, determining the direction of scientific research.

Diversifying the population of researchers is not the only way institutions can mitigate the problem of unconceived alternatives. For example, a bias against the publication of null results disincentivizes research that could limit the generalization of previous findings. Exactly this problem plagued research on social priming, contributing to the replication crisis in psychology. Similar effects may follow from a preference for funding "low-risk" projects by eminent scientists (Stanford, 2019) and the gatekeeping function of prepublication peer review (Heesen & Bright, 2020).

One promising approach is to explicitly incentivize revealing unconceived alternatives. Beyond alignment of verbal and statistical claims, authors could be required to anticipate the unmeasured conditions they suspect are necessary and sufficient to realize the observed effects. This would have the upshot of precommitting researchers to accept certain conceptual replications, even when faced with null results. Building on Yarkoni's advocacy of predictive, translational research, institutions could encourage interdisciplinary collaborations for identifying mechanisms of action and limits to generalization, as well as adversarial collaborations between researchers with competing hypotheses. These efforts could be augmented by the maintenance of prediction markets for aggregating the knowledge and expertise of the broader scientific community (e.g., https://socialsciencepredictionmarket.org/).

These recommendations assume that the causal complexity of social phenomena is not so extreme as to bring researchers to skeptical grief. The (admittedly partial) successes of the social sciences seem to support this assumption, while the field's failures seem plausibly caused by perverse incentives and institutional mismanagement. We are enthusiastic about some of Yarkoni's recommendations – in particular, designing for variation is a fantastic idea and we have sought to do so in our own research (Sievers, Lee, Haslett, & Wheatley, 2019). But insofar as generalization depends on identifying blind spots and missed opportunities, reforms that focus on the statistical practices and thinking habits of individual scientists will likely fall short of the mark. We must also foster a diverse scholarly community that is incentivized to reveal what those who came before them have missed.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Apicella, C. L., Azevedo, E. M., Christakis, N. A., & Fowler, J. H. (2014). Evolutionary origins of the endowment effect: Evidence from hunter-gatherers. *American Economic Review*, *104*(6), 1793–1805. https://doi.org/10.1257/aer.104.6.1793.

Coursey, D. L., Hovis, J. L., & Schulze, W. D. (1987). The disparity between willingness to accept and willingness to pay measures of value. *The Quarterly Journal of Economics*, *102*(3), 679–690. https://doi.org/10.2307/1884223.

Harbaugh, W. T., Krause, K., & Vesterlund, L. (2001). Are adults better behaved than children? Age, experience, and the endowment effect. *Economics Letters*, *70*(2), 175–181. https://doi.org/10.1016/S0165-1765(00)00359-1.

Harding, S. (1992). Rethinking standpoint epistemology: What is "strong objectivity?" *The Centennial Review*, *36*(3), 437–470.

Heesen, R., & Bright, L. K. (2020). Is peer review a good idea? *The British Journal for the Philosophy of Science*, *72*(3), 635–663. https://doi.org/10.1093/bjps/axz029.

Hempel, C. G. (1988). Provisoes: A problem concerning the inferential function of scientific theories. *Erkenntnis*, *28*(2), 147–164.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348. https://doi.org/10.1086/261737.

Knetsch, J. L., & Sinden, J. A. (1984). Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *The Quarterly Journal of Economics*, *99*(3), 507–521.

Lakshminaryanan, V., Keith Chen, M., & Santos, L. R. (2008). Endowment effect in capuchin monkeys. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3837–3844. https://doi.org/10.1098/rstb.2008.0149.

Lange, M. (1993). Natural laws and the problem of provisos. *Erkenntnis*, *38*, 233–248.

List, J. A. (2003). Does market experience eliminate market anomalies?*. *The Quarterly Journal of Economics*, *118*(1), 41–71. https://doi.org/10.1162/00335530360535144.

Maddux, W. W., Yang, H., Falk, C., Adam, H., Adair, W., Endo, Y., … Heine, S. J. (2010). For whom is parting with possessions more painful?: Cultural differences in the endowment effect. *Psychological Science*, *21*(12), 1910–1917. https://doi.org/10.1177/0956797610388818.

Rowe, R. D., D'Arge, R. C., & Brookshire, D. S. (1980). An experiment on the economic value of visibility. *Journal of Environmental Economics and Management*, *7*(1), 1–19. https://doi.org/10.1016/0095-0696(80)90018-2.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*(1), 20–43.

Sievers, B., Lee, C., Haslett, W., & Wheatley, T. (2019). A multi-sensory code for emotional arousal. *Proceedings of the Royal Society B*, *286*(1906).

Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford, UK: Oxford University Press.

Stanford, P. K. (2019). Unconceived alternatives and conservatism in science: The impact of professionalization, peer-review, and Big Science. *Synthese*, *196*(10), 3915–3932.

van Dijk, E., & van Knippenberg, D. (1998). Trading wine: On the endowment effect, loss aversion, and the comparability of consumer goods. *Journal of Economic Psychology*, *19*(4), 485–495. https://doi.org/10.1016/S0167-4870(98)00020-8.

Disentangling paradigm and method can help bring qualitative research to post-positivist psychology and address the generalizability crisis

Moin Syed[a] and Kate C. McLean[b]

[a]Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA and [b]Department of Psychology, Western Washington University (KCM), Bellingham, WA 98225, USA
moin@umn.edu; mcleank2@wwu.edu
https://cla.umn.edu/about/directory/profile/moin;
https://wp.wwu.edu/katemclean/

doi:10.1017/S0140525X21000431, e32

Abstract

For decades, psychological research has heavily favored quantitative over qualitative methods. One reason for this imbalance is the perception that quantitative methods follow from a post-positivist paradigm, which guides mainstream psychology, whereas qualitative methods follow from a constructivist paradigm. However, methods and paradigms are independent, and embracing qualitative methods within mainstream psychology is one way of addressing the generalizability crisis.

Post-positivism, specifically scientific realism, has long been the default philosophical model for psychology (Stedman, Kostelecky, Spalding, & Gagné, 2016). As the field consolidated post-World War II, psychology associated the possibility of scientific credibility via post-positivism with the process of quantification. Rather than critically examining whether quantification is

appropriate for a research question, it became the unquestionable default for the field (Tafreshi, Slaney, & Neufeld, 2016).

Constructivism was a philosophical and methodological response to the pairing of post-positivism with quantification. Qualitative methods, which involve reflexivity and awareness of bias, were a natural fit for the subjectivity central to constructivism. Thus, over time post-positivism became associated with quantitative methods, and constructivism with qualitative methods (Wertz, 2014), and because psychology largely rejected the subjectivity of constructivism, qualitative methods were determined to have no place within mainstream psychology.

What scientific rationale exists to shun qualitative methods? There is none. Research should be driven by well-conceptualized questions that can be addressed empirically (and yes, qualitative data are empirical), but instead mainstream psychology has prioritized the practice of quantification. This practice provides researchers with a sheen of status and opens the door to influence within society. As Yarkoni notes, if psychology is not a quantitative science, will the journalists and policymakers still come knocking? Given the state of our knowledge, perhaps they shouldn't be at all.

Beyond status, a major barrier to embracing qualitative methods in mainstream psychology is the conflation of paradigms and methods (Madill, 2015), which is perpetuated by both post-positivists and constructivists (e.g., Jackson, 2015). Indeed, nearly all discussions of qualitative analysis in psychology are situated within constructivist/critical paradigms (e.g., Gergen, Josselson, & Freeman, 2015). This conflation and divide is so strong that the idea that there might be a place for qualitative methods in psychology is so laughable to the mainstream that Yarkoni had to clearly state that his proposal for greater integration of qualitative methods was sincere. We take Yarkoni's proposal seriously that qualitative methods can address the generalizability crisis, and ask the question, *what would that look like within the post-positivist mainstream of psychology?*

First, we must recognize how qualitative work is already pervasive. Qualitative data that are coded, quantified, and entered into statistical models are common in journals that would otherwise not publish qualitative research (e.g., McLean et al., 2020). Whereas such work may not be perceived as qualitative, per se, it rests on qualitative data and thinking, and highlights how it is not the data source that is the problem but rather the way those data are analyzed.

Yet, in other ways, we have decided as a field that qualitative analysis is just fine. Discussion sections of quantitative articles represent a qualitative analysis of the statistical results, as the authors engage in interpretation and meaning-making, putting their findings in context. As Yarkoni notes, this is the type of inferential procedure that he employed in his arguments. Moreover, measurement studies often involve the identification of latent factors that account for the covariation among indicators. Those latent factors are given names that capture the variation of the indicator set, which is precisely the qualitative analytic process of identifying *themes* (Braun & Clarke, 2006). It appears that even qualitative analysis is permissible in mainstream psychology so long as we do not call too much attention to the practice, and do not engage in the intentionality and rigor of best practices in qualitative methods.

Beyond mere recognition of what we already do, there are two uses of qualitative methods that are underappreciated by post-positivist psychology. Yarkoni proposed that psychology may consider focusing on description, as opposed to the strong emphasis on explanation currently in place. We wholeheartedly agree with this call (Galliher et al., 2017), and add that qualitative methods are particularly well-suited to the task. There is a renewed interest in the critical subject of construct validity (e.g., Grahek, Schaller, & Tackett, 2021), and yet what is often left overlooked is the need to *properly understand* the nature of the construct itself. The ongoing fracas around the construct of ego depletion is an excellent example. Despite hundreds of studies on the topic, amidst the failed replications, it became clear that there was no understanding of what ego depletion even was, let alone how it was related to behavioral outcomes. Some initial qualitative work focused on understanding ego depletion, before moving directly to hypothesis testing, may have saved countless hours (cf., Scheel, Tiokhin, Isager, & Lakens, 2021).

But the role of qualitative methods in psychology should not only be seen as a "first step" that precedes the more central quantitative methods. Qualitative methods can also play a key role in testing, applying, and exemplifying theoretical claims (Robinson & McAdams, 2015). Indeed, Shadish, Cook, and Campbell (2002) distinguished between *causal description* and *causal explanation*, arguing that experiments in psychology primarily address the former. Generating causal explanations is a more formidable task that requires a broader set of methodological approaches, including qualitative methods.

These two uses of qualitative methods, construct development and causal explanations, make clear that qualitative methods have a place within post-positivist psychology, and belie claims that qualitative work is "not science." Such claims stem from the conflation between paradigms and methods, and are more accurately claims about whether or not constructivism is science, which is an argument for another day.

Psychological researchers generally receive no training in philosophy of science or the paradigms and meta-theoretical models of the field. Nor do they generally receive training in qualitative or mixed methods research. Both of these emphasize reflexivity and a focus on the intense complexity of human experience. Psychology has asserted itself as a quantitative field, not through careful study of underlying assumptions and alternatives, but rather through relatively passive absorption of the intergenerational socialization around what it means to conduct *serious science*. This is the great irony of the various ongoing crises within psychology, including Yarkoni's generalizability crisis: that nearly all of the positions and activities that researchers have taken up in the name of serious science are precisely what has exposed our failings.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Galliher, R. V., McLean, K. C., & Syed, M. (2017). An integrated developmental model for studying identity content in context. *Developmental Psychology*, *53*, 2011–2022. https://doi.org/10.1037/dev0000299.

Gergen, K. J., Josselson, R., & Freeman, M. (2015). The promises of qualitative inquiry. *American Psychologist*, *70*(1), 1–9. https://doi.org/10.1037/a0038597

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*(4), 803–815. https://doi.org/10.1177/1745691620966794

Jackson, M. R. (2015). Resistance to qual/quant parity: Why the "paradigm" discussion can't be avoided. *Qualitative Psychology*, *2*(2), 181–198. https://doi.org/10.1037/qup0000031

Madill, A. (2015). Qualitative research is not a paradigm: Commentary on Jackson (2015) and Landrum and Garza (2015). *Qualitative Psychology*, *2*(2), 214–220. https://doi.org/10.1037/qup0000032

McLean, K. C., Syed, M., Pasupathi, M., Adler, J. M., Dunlop, W. L., Drustrup, D., … McCoy, T. P. (2020). The empirical structure of narrative identity: The initial Big Three. *Journal of Personality and Social Psychology*, *119*(4), 920–944. https://doi.org/10.1037/pspp0000247

Robinson, O. C., & McAdams, D. P. (2015). Four functional roles for case studies in emerging adulthood research. *Emerging Adulthood*, *3*(6), 413–420. https://doi.org/10.1177/2167696815592727

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Stedman, J. M., Kostelecky, M., Spalding, T. L., & Gagné, C. (2016). Scientific realism, psychological realism, and Aristotelian–Thomistic realism. *The Journal of Mind and Behavior*, *37*(3/4), 199–218.

Tafreshi, D., Slaney, K. L., & Neufeld, S. D. (2016). Quantification in psychology: Critical analysis of an unreflective practice. *Journal of Theoretical and Philosophical Psychology*, *36*(4), 233–249. https://doi.org/10.1037/teo0000048

Wertz, F. J. (2014). Qualitative inquiry in the history of psychology. *Qualitative Psychology*, *1*(1), 4–16. https://doi.org/10.1037/qup0000007

Mechanistic modeling for the masses

Matthew A. Turner and Paul E. Smaldino

Department of Cognitive and Information Sciences, University of California, Merced, Merced, CA 95343, USA.
mturner8@ucmerced.edu; psmaldino@ucmerced.edu; http://mt.digital, http://smaldino.com

doi:10.1017/S0140525X2100039X, e33

Abstract

The generalizability crisis is compounded, or even partially caused, by a lack of specificity in psychological theories. Expanding the use of mechanistic models among psychologists is therefore important, but faces numerous hurdles. A cultural evolutionary approach can help guide and evaluate interventions to improve modeling efforts in psychology, such as developing standards and implementing them at the institutional level.

Yarkoni says there's a generalizability crisis, and we largely agree. In some ways it's actually worse than he suggests, because of widespread ambiguity and imprecision in specifying theories. It's very hard to test theories if they are imprecise (Smaldino, 2019, 2020). This isn't just a matter of limitations in mapping specific experiments to more general verbal constructs. Rather, those verbal constructs themselves are often so poorly described that severe tests of their applicability become nearly impossible (Mayo, 2018; Popper, 1963). In this light, accounting for more sources of variance in the manner Yarkoni recommends might even be harmful if doing so props up theories that are poorly specified, further insulating such theories from further scrutiny (Smaldino, 2016). It is plausible that failure to clearly specify the components, relationships, and processes in systems of interest leads to exactly those failures to align verbal and statistical models that characterize the generalizability crisis. If one cannot specify how system components influence one another, how could one begin to guess at how observed data might vary? We suggest that this difficulty in precision, which some have named the "theory crisis" (Oberauer & Lewandowsky, 2019), is inherently interlinked with the generalizability crisis. Theories based solely on statistical correlation are notoriously hard to evaluate (Fried, 2020; Meehl, 1990).

Mechanistic explanations and formal models can help by forcing the researcher to articulate their guiding assumptions, decomposing their study system into the parts, properties, and relationships critical for well-formed hypotheses (Kauffman, 1971; Smaldino, 2017, 2020). Mechanistic explanations also allow us to ask "what if things had been different" in a way that non-mechanistic explanations cannot (Craver, 2006). When mechanistic models are operationalized as mathematical or computational models, simulation experiments can be performed that mirror real-world experiments to understand, *a priori*, how certain outcome variables might be affected by treatment or contextual variables (Schank, May, & Joshi, 2014). Even purely verbal mechanistic explanations are a welcome, if marginal, improvement over psychological theories that are too often founded on a series of interesting correlations – each of those correlations potentially a victim of non-replicability or overgeneralization. Mechanistic models facilitate expanded qualitative analyses as well, since mechanistic models can often simplify systems of interest to the point where they can be represented as simple box-and-arrow type diagrams for rapid comprehension, critique, correction, and extension.

Expanding the use of mechanistic explanations in psychology would require a substantial restructuring of the operation and training of psychological scientists. In the long run, this will likely require fairly major institutional change (Smaldino, Turner, & Contreras Kallens, 2019). It is unclear at the present whether the changes will come from within psychology and related departments, or by incursion from other disciplines better trained in formal methods and interested in the juicy problems previously guarded as the domain of psychologists (Smaldino, 2020). Time will tell.

For now, there are some things individuals and institutions can do in the short term to kickstart improvements. Strengthening the theoretical foundations in ways needed to mitigate the generalizability and theory crises requires increased interdisciplinarity, technical expertise, and philosophical scrutiny of assumptions (Smaldino, 2020). Agencies could help by increasing funding for interdisciplinary work between researchers using different approaches to study related problems in the social and behavioral sciences, fostering deeper collaborations between experts in modeling and complex systems and topic experts more familiar with experimental or observational methods. Such projects could also include funding for research software developers and other professional research staff to assist with technical intricacies and provide modeling expertise required to model and analyze complex systems mechanistically. These interdisciplinary teams could collaboratively produce new cyberinfrastructure tools to make mechanistic modeling easier for modeling novices. Later iterations could even extend these tools to automate the generation of computer models and computational experiments to identify sources of variance, and automatically generate statistical models based on the generated computational model (Rand, 2019).

The study of cultural evolution provides some insights into how we might think about the spread of better practices (Gervais, 2021; Smaldino et al., 2019; Smaldino & O'Connor, 2020), including the use and improvement of mechanistic explanations in psychology literature. One approach is to consider that the strategy of failing to align verbal and statistical models is a

communication strategy which has been culturally transmitted to generations of psychology trainees. The lack of specificity and resulting lack of alignment between theory and statistical model may be an instance of deceptive signaling (in the ecological sense, which doesn't imply intent to deceive), where a lack of theoretical rigor is covered up with statistical tests and a recitation of related observed correlations. This maps the generalizability crisis onto analogous problems for which models already exist as starting points, including models of signaling in collaborative environments (Smaldino & Turner, 2020; Smaldino, Flamson, & McElreath, 2018; Tiokhin et al., 2021), the evolution of scientific knowledge on networks (O'Connor & Weatherall, 2018, 2020; Zollman, 2007, 2010, 2013), and the effect of prevailing social power on individual choices (Bergstrom, Foster, & Song, 2016; Henrich & Boyd, 2008; Higginson & Munafò, 2016; O'Connor, 2019). With some further development, these models could be used to conduct several "what if things are different" computational experiments under a variety of assumptions to understand what might happen if various interpersonal or institutional changes were instituted. If any of the considered approaches seem promising *in silica*, it will strengthen the case to expend resources to try them in the real world.

Doing all this is likely to be hard, but worth it. A lack of mechanistic modeling at least complicates the generalizability crisis and is perhaps partly to blame. While this problem is frustrating, it also provides a valuable opportunity to apply social science to an important problem: its own bad state of affairs. The recent, rapid adoption of better practices in psychology and across the sciences, including replication (even if sometimes misguided), registered reports, open science initiatives, data management plans, and more, indicate that many scientists are willing to make changes toward better practices. Changes to institutional incentives must follow.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Bergstrom, C. T., Foster, J. G., & Song, Y. (2016). Why scientists chase big problems: Individual strategy and social optimality. *arXiv*, 1605.05822.
Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.
Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.
Gervais, W. M. (2021). Practical methodological reform needs good theory. *Perspectives on Psychological Science*, *16*(4), 827–843.
Henrich, J., & Boyd, R. (2008). Division of labor, economic specialization, and the evolution of social stratification. *Current Anthropology*, *49*(4), 715–724.
Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, *14*(11), e2000995.
Kauffman, S. A. (1971). Articulation of parts explanation in biology and the rational search for them. In R. C. Buck & R. S. Cohen (Eds.), *PSA 1970* (pp. 257–272). Irvine, CA: Philosophy of Science Association.
Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.
O'Connor, C. (2019). *The origins of unfairness*. Oxford University Press.
O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*(3), 855–875.
O'Connor, C., & Weatherall, J. O. (2020). False beliefs and the social structure of science: Some models and case studies. In D. M. Allen & J. W. Howell (Eds.), *Groupthink in science* (pp. 37–48). Springer.
Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618.
Popper, K. (1963). *Conjectures and refutations*. Routledge.
Rand, W. (2019). Theory-interpretable, data-driven agent-based modeling. In P. K. Davis, A. O'Mahony & J. Pfautz (Eds.), *Social-behavioral modeling for complex systems* (pp. 337–357). Wiley.
Schank, J. C., May, C. J., & Joshi, S. S. (2014). Models as scaffold for understanding. In J. R. Griesemer, W. C. Wimsatt & L. R. Caporael (Eds.), *Developing scaffolds in evolution, culture, and cognition* (pp. 147–167). MIT Press.
Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*, 9.
Smaldino, P. E. (2016). Not even wrong: Imprecision perpetuates the illusion of understanding at the cost of actual understanding. *Behavioral and Brain Sciences*, *39*, e163.
Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, A. Nowak & S. J. Read (Eds.), *Computational social psychology* (pp. 311–331). Routledge.
Smaldino, P. E. (2020). How to build a strong theoretical foundation. *Psychological Inquiry*, *31*(4), 297–301.
Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports*, *8*, 4905. https://doi.org/10.1038/s41598-018-22926-1.
Smaldino, P. E., & O'Connor, C. (2020). Interdisciplinarity can aid the spread of better methods between communities. *MetaArXiv*. Retrieved from https://osf.io/preprints/metaarxiv/cm5v3/.
Smaldino, P. E., & Turner, M. A. (2020). Covert signaling is an adaptive communication strategy in diverse populations. *SocArXiv*. Retrieved from https://osf.io/preprints/socarxiv/j9wyn/.
Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(8), 191249.
Tiokhin, L., Panchanathan, K., Lakens, D., Vazire, S., Morgan, T., & Zollman, K. (2021). Honest signaling in academic publishing. *PLoS ONE*, *16*(2), e0246675.
Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.
Zollman, K. J. S. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass*, *8*(1), 15–27.
Zollman, K. J. S. (2010). Social structure and the effects of conformity. *Synthese*, *172*(3), 317–340.

Generalizability in mixed models: Lessons from corpus linguistics

Freek Van de Velde, Stefano De Pascale and Dirk Speelman

Department of Linguistics, KU Leuven, Blijde Inkomststraat 21/3308, BE-3000 Leuven, Belgium.
freek.vandevelde@kuleuven.be
stefano.depascale@kuleuven.be
dirk.speelman@kuleuven.be
https://www.arts.kuleuven.be/ling/qlvl/people/pages/00039016;
https://www.arts.kuleuven.be/ling/qlvl/people/pages/00102617;
https://www.arts.kuleuven.be/ling/qlvl/people/pages/00013279

doi:10.1017/S0140525X21000236, e34

Abstract

Part of the generalizability issues that haunt controlled lab experiment designs in psychology, and more particularly in psycholinguistics, can be alleviated by adopting corpus linguistic methods. These work with natural data. This advantage comes at a cost: in corpus studies, lexemes and language users can show different kinds of skew. We discuss a number of solutions to bolster the control.

In his assessment of the replicability crisis, Tal Yarkoni points out that the field of psycholinguistics compares relatively favorably with other subdisciplines of psychology. The articles he refers to as commendable advances toward mega-studies (Balota, Yap, Hutchison, & Cortese, 2012; Keuleers & Balota, 2015) have a

classic laboratory design, allowing for multifactorial control over participants and stimuli. Psycholinguistics is, however, not the only field concerned with cognitively plausible accounts of language, nor is it exclusive in its use of quantitatively advanced methods. Usage-based linguistic theories have increasingly turned to large text corpora to answer questions about the cognitive processing of language (Gennari & Macdonald, 2009; Gries, 2005; Grondelaers, Speelman, Drieghe, Brysbaert, & Geeraerts, 2009; Jaeger, 2006; Roland, Elman, & Ferreira, 2006; Piantadosi, Tily, & Gibson, 2011; Pijpops, Speelman, Grondelaers, & Van de Velde, 2018; Szmrecsanyi, 2005; Wiechmann, 2008). Similar to psychology, these studies have steadily turned to generalized linear mixed-effects models to analyze linguistic phenomena (Baayen, 2008; Gries, 2015; Speelman, Heylen, & Geeraerts, 2018).

The advantage of corpus-based studies is that they have higher ecological validity, as they work with naturally occurring data. Additional advantages are (i) the scale of the data, which are usually extracted from corpora that cover millions to even billions of words, reducing the risk of underpowered results; (ii) the high replicability, as the corpora are usually publicly available; and (iii) the possibility to gather data from the past, alleviating the present-day bias to some extent (Bergs & Hoffmann, 2017; De Smet & Van de Velde, 2020; Hundt, Mollin, & Pfenninger, 2017; Petré & Van de Velde, 2018; Wolk, Bresnan, Rosenbach, & Szmrecsanyi, 2013), though the difficulties and obstacles in historical corpus linguistics should not be underestimated (Van de Velde & Peter, 2020). These advantages assuage Yarkoni's concerns about generalizability.

This does not mean that corpus linguistics is a happy-go-lucky picnic. Studies in this field face some daunting difficulties. One is that in corpus data, occurrence frequencies of language users (roughly equivalent to participants) and words (roughly equivalent to stimuli) commonly take a "Zipfian" distribution: word occurrences follow a power law where a few "types" (lemmas) account for most of the "tokens," and most types are in a long tail of infrequent attestations (Zipf, 1935). Similarly for speakers: while observations of a given grammatical construction in a text corpus may come from a wide range of language users (speakers or writers), the distribution is typically skewed such that a few language users contribute a disproportionate amount of the observations. If one wants to use mixed models to investigate the psycholinguistic pressures on the "dative alternance," that is, the difference between *he gave flowers to his mother* versus *he gave his mother flowers*, a heavily investigated phenomenon (see Bresnan, Cueni, Nikitina, & Baayen, 2007; Röthlisberger, Grafmiller, & Szmrecsanyi, 2017 among others), state-of-the-art linguistic corpus studies customarily add a random factor for the verb (*give*, *donate*, *present*, *offer*, *transfer*, *regale*, etc.), but evidently, the corpus will yield many more observations from frequent verbs than from infrequent verbs. If these two factors (words and speakers) are integrated as random factors in mixed-modeling, the maximum likelihood estimation might have a hard time converging on an adequate model: the size of the random intercepts – let alone slopes – may not be reliably estimable with underpopulated levels of the random factors. An often used "solution" is to bin all speakers/writers or word types with less than five observations, but this has the drawback that the underpopulated levels (often the majority) are considered to be the same. This leads to misrepresenting the non-independence of the observations, flouting the very motivation of random effects.

Another problem is that many corpus-based studies suffer from overfitting. This issue is not peculiar to corpus-based studies, but also crops up in other psychological or psycholinguistic studies (Yarkoni, this issue; Yarkoni & Westfall, 2017). The main reason is that corpus linguists tend to use all the data available to fit their mixed model. A solution might come from integrating methods from machine learning (Hastie, Tibshirani, & Friedman, 2013). Repeatedly partitioning the data in training and test sets to carry out cross-validation, bootstrapping, or regularization by shrinkage methods (Ridge, Lasso, and Elastic Net) can reduce the overfit, but at present, applying these techniques in the presence of multiple sources of random variation is not straightforward (see Roberts et al., 2017).

The use of shrinkage methods has an additional application in corpus linguistics, namely when the number of regressors exceeds the number of observations. This could be the case when the lexical effects are focal variables. Instead of treating the different verbs (*give*, *donate*, *present*, *offer*, *transfer*, *regale*, etc.) as the levels of a random factor "verb" when investigating the dative alternance, considering them as merely a source of random variation, we may be interested in their effect on the choice between the two grammatical constructions (… *flowers to his mother* vs. … *his mother flowers*). In corpus linguistics, this is typically achieved by sticking to a verb-as-random-factor approach, focusing on predictions for the random effects, or by running a separate analysis. The former strategy, modeling of focal variables with random factors, arguably "stretches" the purpose of random effects, which are meant to model the association structure in the data, with the fixed-effects modeling systematic trends. The latter strategy often takes the form of "collexeme analysis" (Stefanowitsch & Gries, 2003), but the downside is that it does not work with multifactorial control (Bloem, 2021, p. 115). A promising solution may again come from the aforementioned shrinkage methods (Lasso, Ridge, and Elastic Net) with k-fold cross-validation. K-fold cross-validation is the procedure to repartition the data k times (mostly 10), and each time use 1–1/k of the data as the training set and the remaining 1/k as the test set, in effect iteratively using a small portion of the data as if it were "unseen," to validate the model. Shrinkage with cross-validation not only allows for including a large number of potentially correlating regressors in the model, they also allow for variable selection and effective avoidance of overfitting (Van de Velde & Pijpops, 2019).

Other methodological innovations that are currently explored in linguistics may also contribute to generalizability. An underused technique to check the contours of a statistical model by investigating the effect of the parameters is agent-based modeling. In linguistics, the adoption has been slow, but the last decade has seen an upsurge in such studies (Beuls & Steels, 2013; Bloem, 2021; Landsbergen, Lachlan, Ten Cate, & Verhagen, 2010; Lestrade, 2015; Pijpops, Beuls, & Van de Velde, 2015; Steels, 2016).

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. K. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (pp. 90–115). Psychology Press.

Bergs, A., & Hoffmann, T. (Eds.) (2017). Cognitive approaches to the history of English. Special issue of *English Language and Linguistics*, *21*(2), 191–438.

Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS ONE 8*(3), e58960.
Bloem, J. (2021). *Processing verb clusters*. LOT Dissertation Series.
Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In G. Bouma, I. Krämer, & J. Zwarts (Eds), *Cognitive foundations of interpretation* (pp. 77–96). Amsterdam: KNAW/Edita.
De Smet, I., & Van de Velde, F. 2020. A corpus-based quantitative analysis of twelve centuries of preterite and past participle morphology in Dutch. *Language Variation and Change 32*(3), 241–265.
Gennari, S., & Macdonald, M. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition 111*(1), 1–23.
Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399.
Gries, S. T. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora 10*(1), 95–125.
Grondelaers, S., Speelman, D., Drieghe, D., Brysbaert, M., & Geeraerts, D. (2009). Introducing a new entity into discourse: Comprehension and production evidence for the status of Dutch *er* 'there' as a higher-level expectancy monitor. *Acta Psychologica 130*(2), 153–160.
Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning. Data mining, inference, and prediction* (2nd ed.). Springer.
Hundt, M., Mollin, S., & Pfenninger, S. E. (Eds.). (2017). *The changing English language*. Cambridge: Cambridge University Press.
Jaeger, F. T. (2006). *Redundancy and syntactic reduction in spontaneous speech*. PhD diss., Stanford University.
Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology 68*(8), 1457–1468.
Landsbergen, F., Lachlan, R., Ten Cate, C., & Verhagen, A. (2010). A cultural evolutionary model of patterns in semantic change. *Linguistics 48*(2), 363–390.
Lestrade, S. (2015). A case of cultural evolution: The emergence of morphological case. In B. Köhnlein & J. Audring (Eds.), *Linguistics in the Netherlands* (pp. 105–115). John Benjamins.
Petré, P., & Van de Velde, F. (2018). The real-time dynamics of the individual and the community in grammaticalization. *Language 94*(4), 867–901.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences 108*(9), 3526–3529.
Pijpops, D., Beuls, K., & Van de Velde, F. (2015). The rise of the verbal weak inflection in Germanic. An agent-based model. *Computational Linguistics in the Netherlands Journal 5*, 81–102.
Pijpops, D., Speelman, D., Grondelaers, S., & Van de Velde, F. (2018). Comparing explanations for the complexity principle. Evidence from argument realization. *Language and Cognition 10*(3), 514–543.
Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40*, 913–929.
Roland, D., Elman, J., & Ferreira, V. (2006). Why is 'that'? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition 98*(3), 245–272.
Röthlisberger, M., Grafmiller, J., & Szmrecsanyi, B. (2017). Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics 28*(4), 673–710.
Speelman, D., Heylen, K., & Geeraerts, D. (2018). Introduction. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 1–10). Springer.
Steels, L. (2016). Agent-based models for the emergence and evolution of grammar. *Philosophical Transactions of the Royal Society B 371*, 20150447.
Stefanowitsch, A., & Gries, S.T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8*(2), 209–244.
Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory 1*(1), 113–150.
Van de Velde, F. & Pijpops, D. (2019). Investigating lexical effects in syntax with regularized regression (Lasso). *Journal of Research Design and Statistics in Linguistics and Communication Science*, *6*(2), 166–199.
Van de Velde, F., & Peter, P. 2020. Historical linguistics. In S. Adolphs, & D. Knight (Eds.), *The Routledge handbook of English language and digital humanities* (pp. 328–359). Routledge.
Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory 4* (2), 253–290.
Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica 30*(3), 382–419.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science 12*(6), 1100–1122.
Zipf, G. K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Houghton Mifflin.

Improving the generalizability of infant psychological research: The ManyBabies model

Ingmar Visser[a], Christina Bergmann[b], Krista Byers-Heinlein[c], Rodrigo Dal Ben[c], Wlodzislaw Duch[d], Samuel Forbes[e], Laura Franchin[f], Michael C. Frank[g], Alessandra Geraci[f], J. Kiley Hamlin[h], Zsuzsa Kaldy[i], Louisa Kulke[j], Catherine Laverty[k], Casey Lew-Williams[l], Victoria Mateu[m], Julien Mayor[n], David Moreau[o], Iris Nomikou[p], Tobias Schuwerk[q], Elizabeth A. Simpson[r], Leher Singh[s], Melanie Soderstrom[t], Jessica Sullivan[u], Marion I. van den Heuvel[v], Gert Westermann[w], Yuki Yamada[x], Lorijn Zaadnoordijk[y] and Martin Zettersten[l]

[a]Department of Psychology, University of Amsterdam, Amsterdam, 1018 WB, The Netherlands; [b]Language and Development Department, Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, The Netherlands; [c]Concordia Infant Research Laboratory, Concordia University, Montreal QC H4B 1R6, Canada; [d]Nicolaus Copernicus University, 87-100 Torun, Poland; [e]University of East Anglia, Norwich NR4 7TJ, UK; [f]Department of Psychology and Cognitive Science, University of Trento, 38068 Rovereto, Italy; [g]Stanford University, Stanford, CA 94301 USA; [h]UBC Center for Infant Cognition, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [i]UMass Boston, Baby Lab, Department of Psychology, University of Massachusetts Boston, Boston, MA 02125-3393, USA; [j]Neurocognitive Developmental Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany; [k]School of Psychology, University of Birmingham, B15 2TT Birmingham, UK; [l]Princeton Baby Lab, Princeton University, Princeton, NJ 08540, USA; [m]UCLA Department of Spanish and Portuguese, University of California, Los Angeles, Los Angeles, CA 90095-1532, USA; [n]Department of Psychology, University of Oslo, 0373 Oslo, Norway; [o]Brain Dynamics Lab, University of Auckland, Auckland 1010, New Zealand; [p]Department of Psychology, University of Portsmouth, Portsmouth, UK; [q]Department of Pscyhology, Ludwig-Maximilians-Universität München, 80802 Munich, Germany; [r]Social Cognition Laboratory, University of Miami, Coral Gables, FL 33124, USA; [s]Department of Psychology, National University of Singapore, Singapore 119077; [t]Baby Language Lab, University of Manitoba, Winnipeg, MB R3T 2N2, Canada; [u]Developing Minds Center, Skidmore College, Saratoga Springs, NY 12866, USA; [v]Department of Cognitive Neuropsychology, Tilburg University, 5037 AB Tilburg, The Netherlands; [w]Department: Psychology, Lancaster University, Lancaster LA1 4YW, UK; [x]Kyushu University, Fukuoka, Japan and [y]Trinity College Dublin, Dublin, Ireland
i.visser@uva.nl
christina.bergmann@mpi.nl
k.byers@concordia.ca
dalbenwork@gmail.com
wduch@umk.pl
samuel.forbes@uea.ac.uk
laura.franchin@unitn.it
mcfrank@stanford.edu
alessandra.geraci@unitn.it
kiley.hamlin@psych.ubc.ca
zsuzsa.kaldy@umb.edu
louisa.kulke@fau.de
CML704@student.bham.ac.uk
caseylw@princeton.edu

vmateu@humnet.ucla.edu
julien.mayor@psykologi.uio.no
d.moreau@auckland.ac.nz
iris.nomikou@port.ac.uk
tobias.schuwerk@psy.lmu.de
simpsone@miami.edu
psyls@nus.edu.sg
melsod@babylanguagelab.org
jsulliv1@skidmore.edu
m.i.vdnheuvel@tilburguniversity.edu
g.westermann@lancaster.ac.uk
yamadayuk@gmail.com
l.zaadnoordijk@tcd.ie
martincz@princeton.edu
http://www.ingmar.org
https://www.mpi.nl
https://infantresearch.ca
https://infantresearch.ca
https://www.umk.pl/en
https://people.uea.ac.uk/samuel_forbes
https://webapps.unitn.it/du/en/Persona/PER0169770
http://langcog.stanford.edu
https://webapps.unitn.it/du/en/Persona/PER0033078
https://cic.psych.ubc.ca/
http://babies.umb.edu
https://neurodevpsychology.phil.fau.de/
https://carolinerichards.net/people/
http://babylab.princeton.edu/
https://www.victoriamateu.com/
https://www.sv.uio.no/psi/english/people/aca/julienma/
https://www.braindynamicslab.com
https://www.port.ac.uk/about-us/structure-and-governance/our-people/our-staff/iris-nomikou
https://www.psy.lmu.de/epp/personen/wiss_ma/tobias_schuwerk/
https://people.miami.edu/profile/simpsone@miami.edu
http://blog.nus.edu.sg/lehersingh/
https://babylanguagelab.org/
https://www.skidmore.edu/developing_minds_center/index.php
http://marionvandenheuvel.com
https://www.lancaster.ac.uk/people-profiles/gert-westermann
http://sites.google.com/site/yamadayuk/
https://sites.google.com/view/lorijnzaadnoordijk/homepage
https://babylab.princeton.edu/

doi:10.1017/S0140525X21000455, e35

Abstract

Yarkoni's analysis clearly articulates a number of concerns limiting the generalizability and explanatory power of psychological findings, many of which are compounded in infancy research. ManyBabies addresses these concerns via a radically collaborative, large-scale and open approach to research that is grounded in theory-building, committed to diversification, and focused on understanding sources of variation.

Yarkoni raises concerns about widespread practices in the psychological sciences – ranging from standard statistical practices to narrow experimental designs – which hinder generalizability, theory-building, and ultimately, explanatory power. Infant research in particular faces a range of problems, including difficulties recruiting participants (often resulting in small samples), the unique challenges of designing experiments that hold infants' attention, limited numbers of observations per participant, and infants' rapid developmental changes (Bergmann et al., 2018; Frank et al., 2017; Oakes, 2017).

ManyBabies is a large-scale, multilab collaborative project that currently spans 47 countries and over 200 institutions (https://manybabies.github.io). The project provides a constructive, best-practice, grass-roots approach for addressing issues of replicability and generalizability in infant research and employs a model also utilized by other large-scale, multisite collaborations (e.g., ManyPrimates, 2019; Moshontz et al., 2018). Thus far, ManyBabies has focused its efforts on replicating fundamental findings in infant cognition that underpin our understanding of early cognitive development.

Features and benefits of the ManyBabies approach in addressing the issues Yarkoni identified are (see also Byers-Heinlein et al., 2020; Frank et al., 2017; The ManyBabies Consortium, 2020):

(1) *Consensus-based study designs to advance theory.* ManyBabies projects are focused on evaluating central theories in infant research (e.g., under which circumstances infants show preferences for familiar or novel stimuli in ManyBabies5; Hunter & Ames, 1988), and carefully probing the bounds of theoretical constructs by encouraging participation from researchers with diverse perspectives. ManyBabies' collaborative and consensus-building approach disrupts existing hierarchies, making space for dissent and innovation, and for adjudicating between opposing views (e.g., in the case of adversarial collaboration in ManyBabies2 addressing Theory of Mind; c.f. Baillargeon, Buttelmann, & Southgate, 2018; Cowan et al., 2020; Surian & Geraci, 2012). Simultaneously, it expands collaborative networks to bridge a wide variety of theoretical backgrounds, resulting in designs that clearly identify testable points of disagreement to lay the foundation for further inquiry through experiment and debate.

(2) *Conceptual replications.* As noted by Yarkoni, direct replication is not a sensible target for improving reproducibility if there are concerns about weaknesses in paradigms or stimulus sets that could be addressed in a new experiment (e.g., ManyBabies4 will remove confounds in a paradigm developed to probe infants' social evaluations; Hamlin, Wynn, & Bloom, 2007; Scarf, Imuta, Colombo, & Hayne, 2012). ManyBabies projects probe the generality of phenomena by prioritizing conceptual over exact replications, bringing together researchers from different theoretical and methodological backgrounds to build experimental designs that best capture the processes being studied.

(3) *Diversity in samples and scientists.* By encouraging participation from labs from all over the world and supporting laboratory expenses for scientists who are new to experimental infant research, ManyBabies promotes diversity across multiple dimensions: contexts, lab practices, researchers, and participants. ManyBabies takes seriously the importance and impact of participant heterogeneity (Henrich, Heine, & Norenzayan, 2010), and creates datasets that are more representative of the population of interest (i.e., "human infants") compared to single-lab studies, by testing participants with diverse linguistic and sociocultural backgrounds. Exploring the impact of diversity on the generalizability of core findings has become a prominent target in recent projects, e.g., studying infants at home rather than in a highly-controlled lab setting in ManyBabies-AtHome, thereby reaching more rural populations; assessing the replicability of initial findings with African infants in ManyBabies1A; in ManyBabies3 – studying rule-learning – making the stimuli suitable for infants from different linguistic backgrounds. In doing so, ManyBabies enables us to strike a better balance between

the precision of estimation/breadth of generalization trade-off cited by Yarkoni.

(4) *Quantifying sources of variation*. Studies following the ManyBabies approach can reveal and explicitly measure sources of variation that are difficult to estimate in single-lab studies, including effects of lab practices and methodological variation. For example, ManyBabies1 (addressing infants' preferences for infant-directed speech) tested for effects of distinct experimental methods in infant research (e.g., head-turn preference, central fixation, eye-tracking, ManyBabies Consortium, 2020); ManyBabies2 compares online and in-lab data collection. Both projects thereby probe the generalizability of observed phenomena across experimental paradigms. Specifically, variety is built in through diversity of experimental paradigms used to test a research question – a typical benefit of meta-analysis – yet at the same time we retain control over a number of design factors, as in replication efforts. Given the wide-ranging sources of methodological variation, however, there is considerable work remaining to be done on this issue.

(5) *Stimulus generalizability*. Issues related to stimulus informativeness and generalizability (or lack thereof) are discussed by the ManyBabies project teams and wider community throughout the design process, which generates new "best test" stimuli. The focus is on conceptual replications that involve stimulus sets that differ from the original studies, in this way directly addressing the question of stimulus generalizability. The next step here is to systematically vary stimulus sets.

(6) *Transparent research practices*. ManyBabies is committed to transparency at each research stage, and to collective governance that encourages genuine and non-hierarchical debate, defies the research status-quo, and leads to innovation in theoretical, methodological, and analytic design, as Yarkoni suggests. For example, ManyBabies maintains detailed documentation protocols and openly shares all stimuli and data, including many additional descriptive variables. In this way, additional sources of variance and alternative hypotheses can be tested.

Ensuring that verbal and quantitative expressions of our hypotheses are closely aligned is a tall task. The diversity of scientists involved in each ManyBabies project goes a long way toward developing meaningful operationalizations of the specific research questions under examination. At the same time, the diversity of samples, methods, and stimuli addresses (to an extent) many of the questions on generalizability raised by Yarkoni. Even so, much work remains to tackle concerns related to methodological/stimulus variation, generalizability, and participant heterogeneity, to develop best practices in large-scale international collaborations, and to build better theories (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021). Nevertheless, we look forward to continuing to provide opportunities for learning and growth in the ManyBabies communities, creating the necessary scaffolding for even better research, and, alongside other large collaborative networks, being at the forefront of creating a psychological science that is generalizable and reproducible.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M. B., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. http://doi.org/10.1111/cdev.13079.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647.

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., … Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349. https://doi.org/10.1037/cap0000216.

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., … Logie, R. H. (2020). How do scientific views change? Notes from an extended adversarial collaboration. *Perspectives on Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182.

Hamlin, J., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature 450*, 557–559. https://doi.org/10.1038/nature06288.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 5 pp. 69–95). Ablex.

ManyPrimates. (2019). Collaborative open science as a way to reproducibility and new insights in primate cognition research. *Japanese Psychological Review*, *62*(3), 205–220. https://doi.org/10.24602/sjpr.62.3_205.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186.

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? Simple associations may explain moral reasoning in infants. *PLoS ONE*, *7*(8), e42698. https://doi.org/10.1371/journal.pone.0042698.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*(1), 30–44. https://doi.org/10.1111/j.2044-835X.2011.02046.x.

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809.

An accelerating crisis: Metascience is out-reproducing psychological science

Patrick D. Watson

Minerva Schools at the Keck Graduate Institute, San Francisco, CA 94103, USA. pwatson@minerva.kgi.edu; https://www.patrickdkwatson.com/

doi:10.1017/S0140525X21000121, e36

Abstract

Scientific claims are selected in part for their ability to survive. Scientists can pursue an r-strategy of broad, easy-to-spread ideas, or a K-strategy of stress-tested, bulletproof statements.

The "generalizability crisis" is an exquisite mutation that allows dull, K-strategic methodology articles to spread nearly as quickly as the fast-breeding, r-strategic memes of pop-psychology.

Psychological science has shifted from observational accounts to technology- and statistically-mediated research. Wundt's reaction time experiments describe everyday experience partly because he used simple equipment: a dropped ball-bearing and a telegraph key (Wundt, 1883). Experiments that rely on more complex technologies or analyses require methodological decisions. Because many such decisions are embedded in our research pipelines, our claims are not about directly observed reality. Our data are inextricably connected to the methods used in its collection and analysis. The generalizability (née replicability) crisis stems from the intrinsic trade-off between experimental complexity and generalizability.

For a simple system like Wundt's telegraph key, a small change in input produces a small change in measured output. Such results ought to generalize because we can model noise in the system as Gaussian. But multistep research pipelines are complex systems and a trivial change in inputs can produce huge output changes. The complex pipeline has many hyperparameters and noise in these hyperparameters accumulates with each step of analysis, producing non-Gaussian noise in the output. In complex systems, a more reasonable prior belief is that findings will not generalize.

For example, it is popular, and wrong, to imply that functional magnetic resonance imaging (fMRI) measures brain activity. fMRI measures fluctuations in magnetic fields. These fluctuations are correlated with ionic charge exchange, which in turn is correlated with the deoxygenation of blood, which in turn correlates with brain activity. Because this chain of correlations is so long, it is possible to reject certain claims in the fMRI literature as unrealistically strong (Vul, Harris, Winkielman, & Pashler, 2009). Yet this skepticism of overly strong results also depends on a long chain of correlation: the literature establishing the test-retest reliability of fMRI. Estimates of reliability themselves vary widely depending on the analysis pipeline and experimental parameters. This does not mean that fMRI tells us nothing about the brain, it simply means that the kind of things that fMRI tells us about the brain are interpretable only within a rich context of historical results and consolidated knowledge.

Thus, psychological scientists' linguistic claims about "the mind" are actually claims about the models and methods of their own research contexts. As Yarkoni points out, failure to generalize is a predictable consequence of the gap between the linguistic claims of psychological science and quantitative models. To borrow a statistical term, our discipline has "overfit" its theories to its methods. We've become highly specialized evaluators of our chosen domain. But this specialization precludes a deep understanding of competing research practices; when specialists review others' work, they are only qualified to evaluate surface-level linguistic claims. We are more skillful at predicting the tendencies baked into our own methods, subjects, and analyses than we are at predicting the behavior of people and animals. As we gather more diverse data (Henrich, Heine, & Norenzayan, 2010), we begin to perceive how much our data mirror our own perspectives.

However, just as psychological science is fundamentally interwoven with its own research methods, public scientific debates are best understood in sociohistorical context. All hypotheses struggle for survival. Successful claims must survive a selection process that balances the claims' fertility and lifespan. As in natural selection (Pianka, 1970), scientific "parents" may pursue different strategies to promote the survival of their research "offspring." Yarkoni implies that psychological science is pursuing what is essentially an r-strategy: creating simplistic claims that breed quickly but die off under scrutiny. These claims thrive when audiences are sophisticated enough to engage with linguistic claims, but lack the detailed knowledge required to evaluate the quantitative evidence. Yarkoni argues for a K-strategy: constructing bigger, tougher, longer-lived claims. The replicability community, Yarkoni says, wants durable science built on robust methods.

Ironically, by labeling the entire class of methodological concerns a "crisis," methodologists have invented a broad, qualitative, linguistic claim that is easy to spread. Instead of pedantic discussions of overfit methods, we can point to a lack of rigor with a single phrase. What a brilliant rhetorical move! At a stroke, methodologists can build a broad, interdisciplinary audience with the attention-grabbing, quick-spreading "generalizability crisis," retain their authority as no-nonsense K-strategists, and create a better climate for metascience.

There is ample precedent in the history of science for leveraging manufactured conflict to spread scientific ideas. Even losers reap rewards. We remember Camillo Golgi in the same breath as Ramon y Cajal, even though the former's model of the synapse was flatly incorrect (Glickstein, 2006). More commonly, no one is precisely wrong: In the early 2010s, there was a sometimes contentious debate over whether recollective memory could be clearly distinguished from strong familiarity. This debate fizzled out without clear resolution, yet the conflict produced some fruitful new methods (Koen, Aly, Wang, & Yonelinas, 2013). Even methodological vagueness itself can be a wellspring for debate. Meta-analyses show that the degree of learning transferred between tasks is closely related to the degree of task similarity (Giovanni Sala, Deniz Aksayli, Semir Tatlidil, Tatsumi, & Gondo, 2019). Thus, the "distance" between any pair of tasks could be empirically measured. Yet rather than developing a consensus measure for task similarity, the field actively debates whether a particular finding represents "near" or "far" transfer (Redick et al., 2013). Ambiguous measures give investigators more latitude in building their brand of credulity or skepticism.

Perhaps the generalizability crisis was inevitable. The mind is a system more complex than even the most elaborate research pipeline. All evidence for general principles should be met with extreme skepticism. Reliable quantitative tools can trick us into overgeneralizing when we confuse the tool for the phenomenon. Yet those who can invent more reliable tools have a clear pathway to greater authority within the discipline. As psychological science selects for better tools, it draws practitioners increasingly from the ranks of technologists and statisticians. Our discourse must be tailored to this new social reality. Our audience is us. We are eager for stories where hard-nosed, skeptics skewer weak findings. The move to metascience is a logical endpoint for a discipline studying itself.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Giovanni, S., Deniz Aksayli, N., Semir Tatlidil, K., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology*, *5*(1), 18. doi: https://doi.org/10.1525/collabra.203.

Glickstein, M. (2006). Golgi and Cajal: The neuron doctrine and the 100th anniversary of the 1906 Nobel Prize. *Current Biology*, *16*(5), R147–R151. https://doi.org/10.1016/j.cub.2006.02.053.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83–135. doi:10.1017/S0140525X0999152X. Epub 2010 Jun 15. PMID: 20550733.

Koen, J. D., Aly, M., Wang, W. C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1726.

Pianka, E. (1970). On r- and K-selection. *The American Naturalist*, *104*(940), 592–597. Retrieved February 25, 2021, from http://www.jstor.org/stable/2459020.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359–379. https://doi.org/10.1037/a0029082.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290.

Wundt, W. H. (1883). Ueber psychologische methoden. (On psychological methods). *Philosophische Studien*, *1*, 1–38.

From description to generalization, or there and back again

Kelsey L. West, Kasey C. Soska, Whitney G. Cole, Danyang Han, Justine E. Hoch, Christina M. Hospodar and Brianna E. Kaplan

Department of Psychology, New York University, New York, NY 10003, USA
kelsey.west@nyu.edu
kasey.soska@nyu.edu
wgcole@nyu.edu
danyang.han@nyu.edu
justine.hoch@nyu.edu
christina.hospodar@nyu.edu
brianna.kaplan@nyu.edu

doi:10.1017/S0140525X21000522, e37

Abstract

In his target article, Yarkoni prescribes descriptive research as a potential antidote for the generalizability crisis. In our commentary, we offer four guiding principles for conducting descriptive research that is generalizable and enduring: (1) prioritize context over control; (2) let naturalistic observations contextualize structured tasks; (3) operationalize the target phenomena rigorously and transparently; and (4) attend to individual data.

As developmental researchers, we agree with Yarkoni's assertion that descriptive research offers a potential solution to the generalizability crisis. Careful descriptions of behavior are foundational to psychological science, and especially critical for developmental science where theoretical progress relies on behavior due to children's limited verbal and motor skills. Many scientific fields have rich histories of descriptive work that drive theory building – Galileo's observations of celestial objects, Ramón y Cajal's depictions of neuron structures, and Golgi's cell visualizations. We argue that such descriptions (when done well) are more enduring and valuable than theories based on behaviorally impoverished data. We offer four suggestions to those who want to "take descriptive research more seriously" with examples from developmental science.

Prioritize context over control

Many researchers assume that to understand a psychological phenomenon, they must first distill it into its simplest form. After the fundamentals are established, the idea is that researchers will gradually add in layers of complexity until behavior in the lab resembles natural behavior. However, by prioritizing control over context, researchers may unwittingly sacrifice critical aspects of the original phenomena and risk reifying abstractions that do not generalize beyond a simplified setting. For example, decades of research on the development of walking focused on periodic gait – infants' ability to walk in a straight line over flat ground at a constant speed (see Adolph & Robinson, 2013 for review). Although this simplification enabled researchers to carefully measure infant walking skill, infants rarely walk that way. Instead, at every point in development, infants take omnidirectional steps along curved paths in short activity bursts (Lee, Cole, Golenia, & Adolph, 2018). Work with simulated robots highlights the consequences of "controlling for" these critical aspects of real-world walking. Compared to robots that learned to walk with less variable paths, robots trained with more infant-like, variable paths displayed more functional walking (Ossmy et al., 2018). Thus, prioritizing context over control can help researchers capture the aspects of phenomena that are necessary for generalization.

Let naturalistic observations guide and contextualize structured observations

Researchers' decisions about which methods to use can powerfully shape study outcomes. This is particularly true for infants whose behavior is easily influenced by the environment – who knew, for example, that superfluous sounds can increase infants' attention in looking-time studies? (Spelke, 1985). Indeed, there is power in methods. Caregivers talk far more to their infants during structured play with standardized toys than during daily routines in the home (Tamis-LeMonda, Kuchirko, Luo, & Escobar, 2017). And caregivers' speech is constant during structured play, whereas it ebbs and flows during natural activity. Thus, it is critical for researchers to consider the "facts on the ground" from naturalistic observations as they design, interpret, and generalize data from artificially constructed experimental situations. At minimum, researchers should take care to interpret data from structured tasks as reflecting what infants *can* do – but not necessarily what actually happens in real-world settings.

Rigorously and transparently operationalize behaviors of interest

Researchers should operationalize descriptions of behavior to be robust, straightforward, and transparent. Operational definitions can be tricky. Psychologists typically study higher-order (latent) constructs and may be tempted to quantify constructs by taking a "you-know-it-when-you-see-it" approach, rating the phenomena on an ordinal scale, using yes/no codes, and so on. But gestalt approaches require extensive training to identify constructs reliably and leave future researchers with little information about what participants actually did. Instead, researchers should quantify the actual behaviors. To illustrate, a series of studies documented infants' perception of affordances – whether infants perceive drop-offs and slopes as safe or risky (Adolph & Hoch, 2019). Perception of affordances is a higher-level construct that could be scored as yes or no, but researchers measured it with directly observable

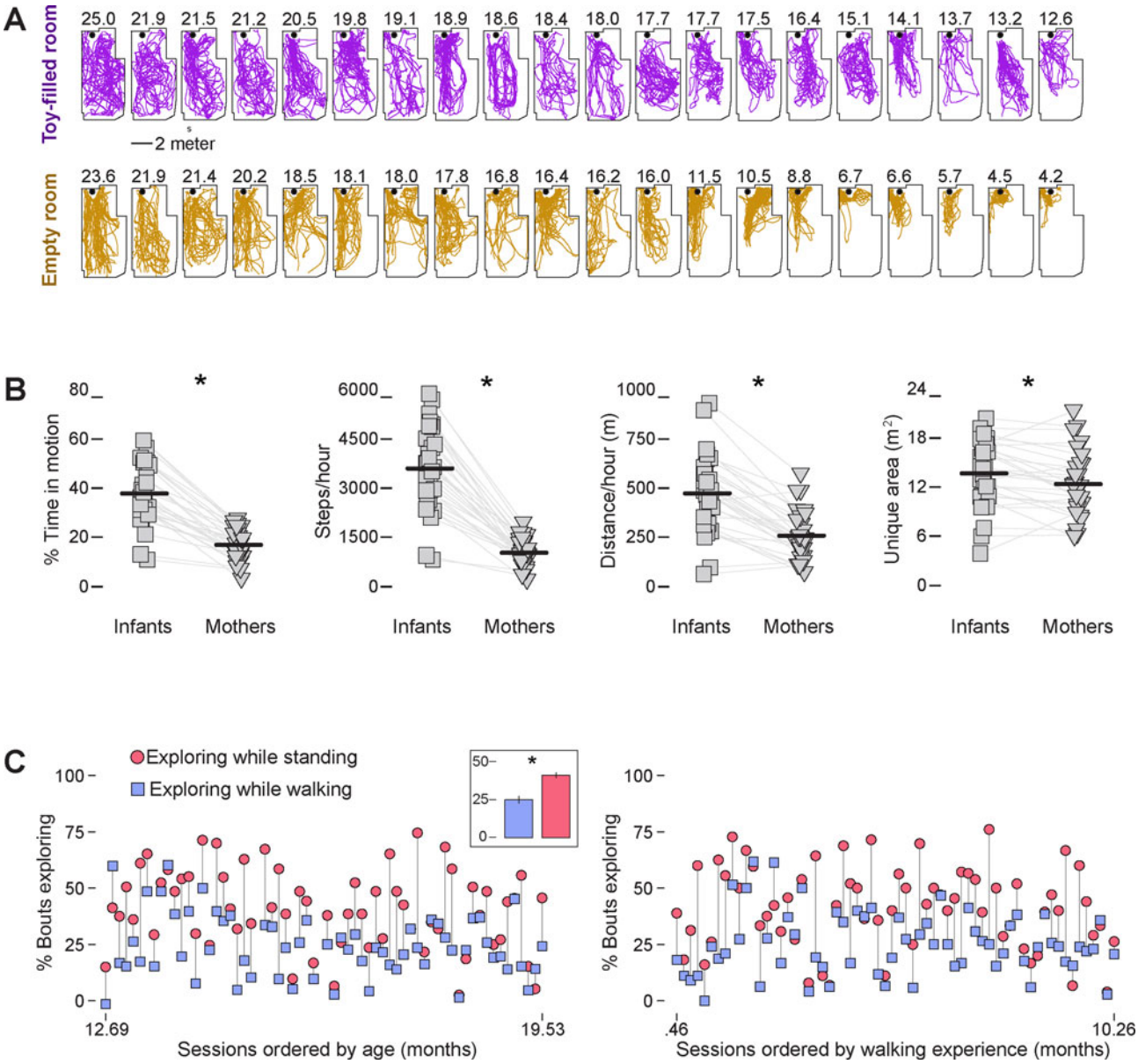

Figure 1. (West et al.) Depictions of individual data that comprise differences between groups or conditions. (A) Infants cover more ground in a toy-filled room than in an empty room. Each plot shows one infant's locomotor path through the toy-filled room (purple) and the empty room (gold) ordered from most to least area covered in m^2. (B) Infants (square symbols) move more than their mothers (triangular symbols) during free play. Gray bars connect each dyad. (C) Infants spontaneously explore objects more frequently while standing (red circles) than while walking (blue symbols) during free play. Each pair of symbols shows one infant's data. Inset shows differences across the group. Infants' propensity to explore objects did not differ by infant age (left panel) or walking experience (right panel).

behaviors such as whether infants attempted to cross, hesitated at the edge, explored the precipice by looking or touching, and displayed negative facial expressions. Such an approach generates a rich description of what happened, including behaviors that may be surprising when considering the abstract construct (e.g., infants rarely display negative emotions when avoiding a risky precipice). Importantly, behavioral descriptions retain their value and will be interpretable to future scientists, whereas higher-level constructs survive only as long as those constructs retain favor.

Attend to individual data

Inter- and intra-individual variability are endemic in development and highly illustrative: Over development, variability can increase or decrease, and the structure of variability can change (Adolph, Cole, & Vereijken, 2015). Thus, ignoring variability can obscure the true nature of phenomena and render generalizations uninformative. Variability is more than measurement error or noise. Rather, understanding each individual's behavior yields better insight into the true nature of the phenomena and can inform mechanisms of change – the nature of the behavior is different if the pattern holds for 95% versus 55% of the sample (Vereijken, 2010). We propose that prior to hypothesis testing with inferential statistics, researchers interrogate each participant's data to assure *themselves* that group-level effects are truly representative. They should use descriptive statistics and simple visualizations to understand the raw data before engaging in complex analyses. Further, to assure *readers* that results are truly representative, plots should show how individual data comprise group differences (e.g., Fig. 1).

Notably, momentum is building in developmental science for large-scale collaborative data collection initiatives, with potential to produce highly generalizable descriptive datasets. Indeed, the Play & Learning Across a Year (PLAY) project leverages 70 labs across North America – with expertise in locomotion, object interaction, emotion, language, gender, environment, and more – to design a common protocol to collect and code videos of 1000+ mothers and infants during natural activity in the home. The data are then shared, so each expert can generate descriptions of behavior that address their own research interests.

As Yarkoni attests, psychology historically focused on testing theories that often fail to generalize to real-world settings. Looking forward, we contend that psychological science should focus on careful, rich descriptions of behavior. Although our suggestions for conducting generalizable descriptive research stem from developmental science, we believe these principles apply broadly across psychological science.

Financial support. Work on this article was supported by NICHD F32 DC017903 to Kelsey West and by NICHD R01 HD-094830, NICHD R01 HD-033486, NICHD R01 HD086034, and DARPA N66001-19-2-4035 to Karen Adolph.

Conflict of interest. None.

References

Adolph, K. E., Cole, W. G., & Vereijken, B. (2015). Intraindividual variability in the development of motor skills in childhood. In M. Diehl, K. Hooker & M. Sliwinski (Eds.), *Handbook of intraindividual variability across the lifespan* (pp. 59–83). New York: Routledge/Taylor & Francis Group.

Adolph, K. E., & Hoch, J. E. (2019). Motor development: Embodied, embedded, enculturated, and enabling. *Annual Review of Psychology*, *70*, 141–164.

Adolph, K. E., & Robinson, S. R. (2013). The road to walking: What learning to walk tells us about development. In P. Zelazo (Ed.), *Oxford handbook of developmental psychology* (pp. 403–443). New York: Oxford University Press.

Lee, D. K., Cole, W. G., Golenia, L., & Adolph, K. E. (2018). The cost of simplifying complex developmental phenomena: A new perspective on learning to walk. *Developmental Science*, *21*, e12615.

Ossmy, O., Hoch, J. E., MacAlpine, P., Hasan, S., Stone, P., & Adolph, K. E. (2018). Variety wins: Soccer-playing robots and infant walking. *Frontiers in Neurorobotics*, *12*, 19.

Spelke, E. (1985). Preferential-looking methods as tools for the study of cognition in infancy. In G. Gottlieb & N. A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 323–363). Norwood, NJ: Ablex.

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., & Escobar, K. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*, *20*, e12456.

Vereijken, B. (2010). The complexity of childhood development: Variability in perspective. *Physical Therapy*, *90*, 1850–1859. doi: 10.2522/ptj.20100019.

Generalizability challenges in applied psychological and organizational research and practice

Brenton M. Wiernik, Mukhunth Raghavan, Tyler Allan and Alex J. Denison

Department of Psychology, University of South Florida, Tampa, FL 33620, USA
wiernik@usf.edu, mukhunth@usf.edu, tylera1@usf.edu, adenison@usf.edu

doi:10.1017/S0140525X21000492, e38

Abstract

Yarkoni highlights patterns of overgeneralization in psychology research. In this comment, we note that such challenges also pertain to applied psychological and organizational research and practice. We use two examples – cross-cultural generalizability and implicit bias training – to illustrate common practices of overgeneralization from narrow research samples to broader operational populations. We conclude with recommendations for research and practice.

Body

Yarkoni's critique focuses on overgeneralization from narrow sets of experimental stimuli, situations, or manipulations to very broad theoretical constructs. The issue of generalizability is not limited to academic and theoretical work but is also critically relevant for applied work that informs decisions made by organizations, institutions, and governments. Historically, applied psychology suffered from undergeneralization – a mistaken belief in "situational specificity," that effects were unique to specific contexts and never (or rarely) generalizable (Schmidt & Hunter, 1977). With the advent of meta-analysis, applied psychology learned that much of this apparent variability was due to statistical artifacts. However, the field may have overcorrected and now tends to overgeneralize. Frequently, models, measures, or interventions are "validated" in narrow settings, then applied in other contexts without carefully considering generalizability. Where generalizability is tested, it is often done in limited ways. In this commentary, use two key generalizability challenges as illustrations.

First, consider cross-cultural generalizability. Much applied psychology research occurs in the United States and Western Europe, but these models are commonly used to motivate research and inform operational practices (e.g., selection systems, assessments, interventions) in organizations around the world (Gelfand, Leslie, & Fehr, 2008). Researchers may allude to potential cross-cultural differences, but the core of Western models is generally assumed to apply across cultural contexts. Where generalizability across cultures is attended to, it is often done so haphazardly. A study might compare results in samples drawn from only two countries, rather than from a wider range of countries (Ones et al., 2012). Often, these countries are described as varying on a single dimension of cultural characteristics (e.g., collectivistic vs. individualistic). Despite the narrow sampling of cultures in these studies, broad conclusions are often drawn about "collectivism" and assumed to apply to any culture that could be classified into these categories. Finally, samples in studies are frequently drawn from narrow subgroups within a culture (e.g., university students) without consideration of how these groups may differ from other cultural groups within a country. Several common overgeneralizations are apparent:

1. Individual countries are assumed to be exchangeable with others similarly classified.
2. The focal cultural characteristics (e.g., collectivism) are assumed to be the operative cause of cultural effects rather than other unmodeled factors (e.g., power distance, religiosity, history, economic environment).
3. The specific populations or subcultures sampled in a country are assumed to be representative of a country's cultural diversity.

These failures to consider included countries, characteristics, and subgroups as sampled from broader populations of these entities limit what conclusions are justifiable. If a researcher or organization observes that an intervention is effective in two contexts, they may conclude it will be effective more broadly. Conversely, if a study observes differences in relationships between two settings, they may also *overestimate* the variability in the relationship across cultures more broadly.

To better justify cross-cultural generalizations, researchers and practitioners must consider how representative their samples of individuals, characteristics, and countries are broadly. One possible approach is to conduct studies broadly sampling from diverse cultures around the world (Ones et al., 2012). By robustly sampling from many cultures, researchers can more accurately gauge whether relationships are consistent or variable across contexts. An alternative approach would be to extend conclusions about generalizability cautiously. Conclusions from narrow studies should be limited to the groups, countries, and contexts represented. In reports, investigators should conclude "X predicts Y in a sample of students in the United States" rather than making generalizations about broad cultural factors such as "collectivism." Over time, as single-context samples accumulate, systematic reviews of this evidence can identify the patterns that generalize (Oh, 2009; van Aarde, Meiring, & Wiernik, 2017). Such an approach encourages appropriate caution and also encourages de-centering of Western perspectives, allowing researchers themselves representing diverse cultures to pose questions that are relevant to their cultural contexts (Cheung, van de Vijver, & Leong, 2011; Gelfand et al., 2008).

Second, consider validation of interventions. Organizational interventions are frequently trialed in narrow populations (e.g., university students or employees from a few organizations), then deployed operationally without careful evaluation of their broader effectiveness. A recent popular example is implicit bias-based interventions to address issues of bias, racism, and inequities in organizations. Such interventions have become widespread, but evidence for their effectiveness for improving prejudice and inequity outcomes is sparse (FitzGerald, Martin, Berner, & Hurst, 2019; Onyeador, Hudson, & Lewis, 2021). In a systematic review of implicit bias interventions, Forscher et al. (2019) found that the large majority of studies were conducted with US university students, focused only on changes in implicit attitudes versus broader outcomes, and reported small effects. Importantly, Forscher et al. observed substantial heterogeneity across studies, underscoring that broad generalizability of implicit bias effects should not be expected. In light of this review, the uptake of implicit bias interventions for operational use has outpaced evidence supporting them. We identify three key areas of overgeneralization:

1. Generalization from studied populations (primarily university students) to operationally-relevant population (employees in specific industries and regions).
2. Generalization from studied outcomes (primarily short-term changes in implicit bias scores) to operationally-relevant outcomes (prejudice and equity outcomes).
3. Generalization from small observed effects to assume larger societal relevance. Despite observing small effects, studies frequently allude to potential large societal impacts (e.g., through accumulation across people or over time; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2015).

These overgeneralizations have consequences. Not only may these interventions be ineffective, but they appear to crowd out other actions that may better address systemic inequities (Pritlove, Juando-Prats, Ala-leppilampi, & Parsons, 2019). Organizational equity, diversity, and inclusion efforts should adapt to emphasize practices with stronger, more generalizable evidence bases such as intergroup contact and systems to bypass individual prejudice (Onyeador et al., 2021).

We highlight these two examples as part of a broader pattern of overgeneralization in applied psychology from narrow samples, contexts, and measures to broader constructs and populations. To ensure effectiveness of organizational practices, we urge applied researchers and practitioners to make generalizations more cautiously. In particular, we urge organizations to await evidence on operationally-relevant groups and measures (e.g., actual diversity and equity outcomes) before moving models and interventions into practice.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *66*(7), 593–603. https://doi.org/10.1037/a0022389

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology*, *7*(1), 29. https://doi.org/10.1186/s40359-019-0299-7

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Gelfand, M. J., Leslie, L. M., & Fehr, R. (2008). To prosper, organizational psychology should… adopt a global perspective. *Journal of Organizational Behavior*, *29*(4), 493–517. https://doi.org/10.1002/job.530

Oh, I.-S. (2009). *The Five Factor Model of personality and job performance in East Asia: A cross-cultural validity generalization study*. Doctoral dissertation, University of Iowa. http://search.proquest.com/dissertations/docview/304903943/

Ones, D. S., Dilchert, S., Deller, J., Albrecht, A.-G., Duehr, E. E., & Paulus, F. M. (2012). Cross-cultural generalization: Using meta-analysis to test hypotheses about cultural variability. In A. M. Ryan, F. T. L. Leong & F. L. Oswald (Eds.), *Conducting multinational research projects in organizational psychology: Challenges and opportunities* (pp. 91–122). American Psychological Association. https://doi.org/10/5kz

Onyeador, I. N., Hudson, S. T. J., & Lewis, N. A. (2021). Moving beyond implicit bias training: Policy insights for increasing organizational diversity. *Policy Insights from the Behavioral and Brain Sciences*, *8*(1), 19–26. https://doi.org/10.1177/2372732220983840

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*(4), 562–571. https://doi.org/10.1037/pspa0000023

Pritlove, C., Juando-Prats, C., Ala-leppilampi, K., & Parsons, J. A. (2019). The good, the bad, and the ugly of implicit bias. *The Lancet*, *393*(10171), 502–504. https://doi.org/10.1016/S0140-6736(18)32267-0

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529–540. https://doi.org/10.1037/0021-9010.62.5.529

van Aarde, N., Meiring, D., & Wiernik, B. M. (2017). The validity of the Big Five personality traits for job performance: Meta-analyses of South African studies. *International Journal of Selection and Assessment*, *25*(3), 223–239. https://doi.org/10.1111/ijsa.12175

The stimulus-response crisis

Robyn Wilford[a], Juan Ardila-Cifuentes[b], Edward Baggs[c] and Michael L. Anderson[b,c,d]

[a]Department of Psychology, University of Western Ontario, London, ON N6A 3K7, Canada; [b]Department of Philosophy, University of Western Ontario, London, ON N6A 3K7, Canada; [c]The Rotman Institute of Philosophy, University of Western Ontario, London, ON N6A 3K7, Canada and [d]The Brain and Mind Institute, The University of Western Ontario, London, ON N6A 3K7, Canada.
rwilfor@uwo.ca; jardilac@uwo.ca; ebaggs@uwo.ca; mande54@uwo.ca; www.emrglab.org

doi:10.1017/S0140525X21000285, e39

Abstract

Yarkoni correctly recognizes that one reason for psychology's generalizability crisis is the failure to account for variance within experiments. We argue that this problem, and the generalizability crisis broadly, is a necessary consequence of the stimulus-response paradigm widely used in psychology research. We point to another methodology, perturbation experiments, as a remedy that is not vulnerable to the same problems.

Although Yarkoni frames his primary concern with current psychology research paradigms in terms of the frequent mismatch between verbal and statistical expressions of their hypotheses, the main problem he uncovers is the failure to account for, measure, or control variance. Perhaps most important is the common failure to account for variation in the experimental materials themselves – the "stimulus as fixed effect fallacy." Here we would like to point out that these problems, which Yarkoni correctly identifies, are a necessary consequence of psychologists' abiding commitment to the stimulus-response (S-R) formula when constructing experiments. The assumptions behind this style of psychological investigation are the root cause of the operationalization and generalizability crisis.

The S-R formula assumes that you can reduce a psychological phenomenon to a simplified behavioral response to an isolated perceptual cue. For example, "attention" is measured by recording response time to select letter presentations in the presence or absence of distractors, or, in Yarkoni's example, "recognition memory" is measured by asking participants to select a target photograph of a previously-seen face under different dual-task interference conditions. This kind of reductive operationalization of the psychological phenomenon appears to offer experimental control, but in fact hides real subject-induced variance (Dewey, 1896), removes the phenomenon from the actual contexts in which it manifests (Danziger, 1994, pp. 30–33), and in some cases may well destroy the phenomenon entirely (Gibson, 1979, pp. 1–4). All of these are necessary consequences of the S-R model and lead directly to problems of generalizability. There is, however, a different way to proceed. As an additional remedy to the issues Yarkoni raises, we would like to draw attention to a class of methods, sometimes called perturbation experiments, that approach the study of perception and behavior differently.

A traditional S-R experiment asks questions of the form, "if I present this isolated cue to a participant, what response is elicited?" The aim is to establish a link in statistical terms between the thing being experimentally varied (the "stimulus," or the independent variable) and the behavior being measured (the "response," or the dependent variable). Because the question is answered through these statistical means, and finding an experimental effect depends vitally on controlling variability within the experiment other than the intended experimental manipulation, these S-R experiments are necessarily vulnerable to the failure of accounting for sources of variance, a problem that, as Yarkoni shows, can quickly become intractable.

In a perturbation experiment, the aim is different. Perturbation experiments aim to identify the precise variable or variables implicated in the ongoing control of a complete activity. A perturbation experiment asks questions of the form, "precisely which aspects of this ongoing activity do I need to disrupt in order to cause a qualitative shift in the behavior?" This kind of methodology has a long history in physiological psychology. Classic nineteenth century studies of brain injury are a form of natural perturbation experiment (Damasio et al., 1994; James, 1890, Ch. 2). Modern transcranial magnetic stimulation studies in human subjects, and optogenetic methods in animal models, are perturbation experiments in which a physiological perturbation is introduced artificially by the experimenter.

Perturbation methods have long been used in behavioral studies too, notably in motor control studies (e.g., Gibson & Walk, 1960). We would like to draw attention to their use in a motor development study looking at how infants negotiate slopes of varying inclination (Adolph, Eppler, & Gibson, 1993). This study found that, while crawling infants attempt to descend too-steep slopes head-first, older, more experienced toddlers modify their style of locomotion before attempting the descent (e.g., sliding down instead of attempting to walk down). The qualitative bifurcation in the behavior of the toddlers – the slope's perturbation of their default mode of locomotion – is unambiguous evidence of their having learned to attend to the visual cue for slope. Conclusions drawn from perturbation experiments do not depend on establishing links between the phenomenon and what causes it in statistical terms, and so they are not vulnerable to the stimulus as fixed-effect fallacy and other failures in accounting for variance. Further, because this methodology allows complete and ongoing behavior, it keeps the phenomenon of interest and the context in which it happens relatively intact. Note, as well, in contrast to typical S-R perceptual cues, the perceptual cue in this case (the inclination of the slope) is not isolated nor it is fixed; and, by systematically changing this variable, these experiments do not sacrifice methodological rigor.

A challenge is how to scale the perturbation methodology from investigation of "online" motor control tasks to more cognitive tasks, such as those in the experiments Yarkoni discusses putatively demonstrating the verbal overshadowing effect. In response, we would make two points. First, certain kinds of cognitive abilities are more immediately amenable to the perturbation paradigm than others. Decision-making and attention may be relatively amenable to perturbation methods. In the Adolph et al. (1993) study, the bifurcation in the toddlers' behavior when the slope becomes too steep is evidence that they have learned to *attend* to the visual cue for the slope (which necessarily means they perceive it), and as a result they have *decided* to locomote in a different way. The study can be interpreted as measuring attention and decision-making *in situ*. Of course, more work needs to be done to extend this methodological approach to higher-order symbolic forms of cognition (Baggs, Raja, & Anderson, 2020).

Second, by assuming that the only way to establish a psychological fact is via experiments set up in the S-R format, psychology unnecessarily constrains itself. The S-R methodology has been the dominant method used in psychology labs for 150 years, and this situation has led repeatedly to periods of crisis, of which the current version – focused on replicability and now on generalizability – is merely the latest iteration (Reed, 1996, pp. 3–5). Perhaps it is

time to recognize that there are more methods available in the psychologist's toolbox than what is dreamt under the S-R philosophy.

Financial support. This work was supported by a Canada Research Chair award to MLA (award # 950-231929 from SSHRC). The authors declare they have no conflicts of interest pertaining to the material presented here.

References

Adolph, K. E., Eppler, M. A., & Gibson, E. J. (1993). Crawling versus walking infants' perception of affordances for locomotion over sloping surfaces. *Child Development*, *64*(4), 1158–1174.

Baggs, E., Raja, V., & Anderson, M. L. (2020). Extended skill learning. *Frontiers in Psychology*, *11*, 1956. https://doi.org/10.3389/fpsyg.2020.01956.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science*, *264*(5162), 1102–1105.

Danziger, K. (1994). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*(4), 357.

Gibson, E. J., & Walk, R. D. (1960). The "visual cliff." *Scientific American*, *202*(4), 64–71.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

James, W. (1890). *The principles of psychology* (Vol. *1*). Henry Holt.

Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology*. Oxford University Press.

Author's Response

Replies to commentaries on the generalizability crisis[1]

Tal Yarkoni

Department of Psychology, The University of Texas at Austin, Austin, TX 78712-1043, USA
tyarkoni@gmail.com

doi:10.1017/S0140525X21001758, e40

Abstract

The 38 commentaries on the target article span a broad range of disciplines and perspectives. I have organized my response to the commentaries around three broad questions: First, how serious are the problems discussed in the target article? Second, are there are other, potentially more productive, ways to think about the issues that the target article framed in terms of generalizability? And third, what, if anything, should we collectively do about these problems?

R1. Overview

The 38 commentaries on the target article span a broad range of disciplines and perspectives. I have organized my response to the commentaries around three broad questions: First, how serious are the problems discussed in the target article? Second, are there are other, potentially more productive, ways to think about the issues that the target article framed in terms of generalizability? And third, what, if anything, should we collectively do about these problems? Each of these three sections is, in turn, divided into a number of subsections, each of which summarizes a particular answer to the question given by one or more commentaries.

The assignment of commentaries to groups is necessarily loose. Also, several commentaries show up in multiple groups, because, despite being under 1,000 words each, they are large; they contain multitudes.

R2. How serious is the problem?

A sensible place to start a review of 38 commentaries on a fairly polemical article is to ask to what extent those commentaries agreed or disagreed with the article's general characterization of affairs. The target article's central claim was that psychology research presently suffers from widespread failure to adequately align verbal hypotheses with their presumed statistical operationalizations, resulting in pervasive generalizability failures: Researchers often know very little about what universe of observations their statistical results actually refer to, and consequently draw overly broad conclusions that the statistics do not support on any obvious reading. Although the great majority of commentaries expressed substantive agreement with this claim, several did not – and even among those that agreed, there were differences in commentators' positions. Here, I characterize four different positions, ranging from outright rejection of the target article's central premise to wholesale acceptance of the argument and exploration of some of the more severe consequences.

R2.1. No big problems

Two commentaries argue that the generalizability-related problems highlighted in the target article focuses are already widely appreciated and have well-established solutions. The stronger position is taken by **Lakens, Uygun Tunç, and Tunç (Lakens et al.)**,[2] who argue that researchers already have two perfectly sound ways to justify generalizability claims – falsificationism and confirmationism. In Lakens et al.'s view, the argument laid out in the target article constitutes "a third approach built on the impossible ideal of verifying (i.e., conclusively confirming) generalizability claims through random-effect modelling." To be frank, I am not sure how the authors arrive at this conclusion. So far as I can see, my paper says nothing that could be reasonably construed as a new philosophy of science. It simply points out the direct implication of what seems to me an incontrovertible fact: One cannot pair up verbal claims with statistical models arbitrarily and still claim one is doing science. There is no serious philosophical view under which one may freely draw whatever verbal conclusion one wishes to from a given statistic, irrespective of the latter's consensual meaning. Neither falsificationists nor confirmationists get to pretend otherwise. For the falsificationist, a deductively sound conclusion requires true premises – and how could the truth of a premise like "my theory is falsified if I observe that $A > B$, $p < 0.05$" not depend on the specification of the model that produced the p value? Correspondingly, how could a confirmationist ever conclude, as Lakens et al. suggest, that "subsequent observations enlarge the set of positive instances predicted by the theory" unless the confirmationist understands what set of instances a given statistical model plausibly refers to? When an author observes $A > B$ in a particular experiment, should they write in their Discussion that the effect is present for all possible populations, for only the specific observations in the sample, or for some intermediate universe in between? And, just how do Lakens et al. think any researcher – be they a falsificationist,

confirmationist, or anarchist – could make such a determination without thinking carefully about their model specification?

Gilead argues that the generalizability of a finding is often only a secondary concern, as researchers often operate in other modes of investigation. There are two ways to read this concern. One reading is that Gilead is arguing that psychologists don't always need to lean on inferential statistics so heavily; that when they are doing what Gilead calls *naming* or *causal ontology*, they can rely on other methods of inference. If this is Gilead's point, I agree with it – indeed, several of my recommendations are to exactly this effect. But there is another reading under which Gilead is saying something much stronger – something more like, *hey, lighten up – the way people use inferential quantities like p-values is fine, even if those quantities don't map onto reality in quite the way the authors' words seem to suggest*. The latter view seems implied by Gilead's assertion that "the generalizability of a pattern is fully independent from the claims made about its generalizability" – by which he means, I think, that it does not matter much what verbal conclusions authors draw from their statistical results, because readers are always free to draw their own conclusions, and so "it is irrelevant whether an author is grandstanding." If *this* is the intended force of Gilead's argument, it seems to me clearly wrong. The point of writing scientific papers is, I think, to clearly and accurately communicate one's findings to others. The fact that a very motivated reader with enough expertise and free time on their hands could, in principle, carefully pore over the methods and results of every paper they read before drawing their own conclusions doesn't seem like a good reason to ignore the strong claims authors routinely make in their manuscripts.

R2.2. Big problems, but no crisis

Two of the commentaries – **Medaglia and Fernandez** and **Watson** – acknowledge the severity of the problems I draw attention to, but argue that they don't rise to the level of a crisis. The concern, as Medaglia and Fernandez express it, is that "[t]he recent trend to label dilemmas in psychology as 'crises' is insidious," and risks "contributing to bandwagoneering negativity, cynicism, indifference, and antiscientific sentiments." I am sympathetic to this argument in principle, inasmuch as one can clearly cause harm by exaggerating the implications of a situation – that is, after all, one of the central claims of the target article! But the crisis label seems to me wholly appropriate in this case. Medaglia and Fernandez do not dispute the arguments made in the paper; on the contrary, they explicitly endorse them. Yet the direct implication of these arguments is that psychologists routinely make claims that are not only spurious on a reasonable reading of the marshaled statistics, but also often have no meaningful connection to the empirical data at all. If this doesn't constitute a crisis for a field, personally, I have a hard time imagining what would.

R2.3. Big problems, but…

Several commentaries agree with the general tenor of the target article, but in a qualified way: They argue that the problems discussed in the target article are downstream symptoms of some more fundamental failing, and that the situation is unlikely to improve much until the root cause (whatever it may be) is addressed. I discuss most of these commentaries in more detail in section R.3, as they generally include some argument to the effect that the so-called generalizability crisis is better understood or conceptualized in different terms. **Dacey**'s commentary is, perhaps, unique in that the author embraces my characterization of the problems but nevertheless argues that many of them could be readily eliminated if researchers were to simply "recognize the distinction in statistics between statistical hypotheses and substantive hypotheses, and to treat them differently from one another." Although I don't exactly disagree with this suggestion, I'm not sure what it adds to the analysis. It goes without saying that substantive and statistical expressions are not the same thing, and should be treated separately; indeed, if they *were* the same thing, there would have been no point in the first place in my arguing that researchers should take greater pains to align the two. Therefore, what Dacey sees as a solution is to my mind simply a restatement of one of the target article's central premises.

Three other commentaries [**Braver & Braver, Sievers & DeFilippis**, and **Iliev, Medin, & Bang (Iliev et al.**)] argue that the target article is, despite the soundness of most of its arguments, too pessimistic in outlook. Braver and Braver don't like my suggestion that some psychologists may wish to consider a different career or focus more heavily on qualitative research. Sievers and DeFilippis argue that the target article makes much of social science sound hopelessly difficult, and suggest that such pessimism is unwarranted given that there are numerous examples of robust psychology findings in the literature. Iliev et al. only mention that my outlook is "gloomy" in passing. Braver and Braver, presumably, do not mean to suggest that it is *never* appropriate for researchers to question whether they could or should be doing something working in the public interest *not* to occasionally ask themselves whether what they're doing is worthwhile. The same logic applies when determining whether a given research problem is or is not tractable. Sievers and DeFilippis would surely agree that the mere fact that many psychology findings are robust does not mean that *every* research question that pops into one's head must be worth pursuing. The point is, it falls on each individual researcher to ask whether *their* particular question seems likely to yield fruit. If the answer promotes gloom, so be it.

R2.4. Big problems, and…

The final, and largest, subgroup consists of commentaries that accept the target article's central premises more or less as-is, and focus their discussion either on potential solutions to the problems or on further exploration of the implications. I discuss the former set of commentaries in section R4; here, I focus on the latter – that is, those commentaries that expand on the issues raised in my article. Several of these commentaries draw attention to the implications for applied research: **Grubbs** focuses on clinical psychology applications; **Wiernik, Raghavan, Allan, and Denison** focus on issues in industrial-organizational settings; **de Leeuw, Motz, Fyfe, Carvalho, and Goldstone** focus on implications for education; and **Brewin** discusses implications in the legal sphere. I found all of these commentaries lucid and compelling, but lack of expertise in these areas precludes me from adding much of substance. The shared message of these commentaries – namely, a recognition that the rampant overgeneralization common to many areas of psychology is not just a distasteful but benign consequence of systemic pressures and warped incentives, but can and does routinely lead practitioners and policy-makers to deploy suboptimal and even dangerous real-world interventions. Grubbs makes probably the strongest claim in this respect – although one that I think is entirely justified – when he points

out that, in clinical psychology (although the point applies to many other applied fields), "the costs of a generalizability crisis are measured in human lives, not wasted resources'."

Two commentaries focus on implications for basic research in specific domains of psychology. **Visser, Bergmann, Byers-Heinlein, Dal Ben, Duch, Forbes, Franchin, Frank, Geraci, Hamlin, Kaldy, Kulke, Laverty, Lew-Williams, Mateu, Mayor, Moreau, Nomikou, Schuwerk, Simpson, Singh, Soderstrom, Sullivan, van den Heuvel, Westermann, Yamada, Zaadnoordijk, and Zettersten (Visser et al.)** describe key features of the ManyBabies initiative (Frank et al., 2017), and illustrate how these can help address many of the problems described in the target article, which is a very reasonable approach. **Harris, Pärnamets, Brady, Robertson, and Van Bavel (Harris et al.)** discuss implications for moral and political psychology.

Finally, several commentaries suggest that the target article may, actually, have *under*stated the severity of the problems it describes. **Turner and Smaldino** point out that theories in psychology are often so underspecified as to be essentially untestable – in which case, what does it even matter which variance components are or are not included in a model? **Gelman** observes that the problems I discuss are not limited to psychology, and also pervade many other sciences. The latter point is also echoed by **Maniadis**, who notes that similar troubles afflict experimental economics, a field that is (at least on its face) far more quantitatively rigorous than most of psychology.

R3. Other ways to conceptualize the problem

R3.1. Lack of theory

Several commentaries view the root problem underlying the issues the target article describes as a lack of adequate theory. Appeals for more theory in psychology are, of course, an old phenomenon – although they do seem to be experiencing something of a renaissance recently, Many, however, never bother to tell us what they actually mean by *theory*. Several of the present commentaries [e.g., **Maniadis**; **Harris et al.**; **Visser et al.**; **Lakens et al.**; **Davidson, Ellis, Stachl, Taylor, & Joinson (Davidson et al.)**; and **Turner & Smaldino**] fall into this category. Most of these commentaries call for more theory only tangentially, so it is, perhaps, unfair to expect a detailed explication. But, in a couple of cases, lack of theory is the primary focus of the commentary, and still the reader is given no clear definition. For example, **Hensel, Miłkowski, and Nowakowski**'s (**Hensel et al.**) titular claim is that "Without more theory, psychology will be a headless rider." The reader is never, actually, told what Hensel et al. *mean* by theory, but the definition must be an inclusive one indeed, for the authors take pains to note that, despite the absence of any explicit discussion of theory in the target article, "[Yarkoni] wouldn't have been able to make his case without appealing to theoretical insights."

Let us suppose that **Hensel et al.** are right. Well, what then? So far as I can see, my argument relies almost entirely on a bit of statistics and some common sense, making no appeal to any domain expertise in psychology. Perhaps, all the authors really mean by *theory* is, roughly, *careful thinking*, a difficult position to argue against. The trouble is, Hensel et al. don't tell us how to differentiate good theory from bad theory, or how a bad theorist might go about becoming a better one. "Theorizing," we are simply told, "is an activity integral to any scientific approach regardless of its specific aims and methods. It transcends the difference between qualitative and quantitative research." But when theory is everything it is also nothing. No surprise if Hensel et al. believe that "all the shortcomings of current practice discussed by Yarkoni come from a common source: researchers' inadequate appreciation of how various theoretical considerations should inform the decisions made at every stage of scientific investigation." It could hardly be otherwise.

A similar concern applies to **Turner and Smaldino**'s call for greater use of mechanistic models in psychology. Here, again, the conclusion that psychologists should use more mechanistic models is not clear about what concrete approaches the authors are actually advocating. We are told that mechanistic explanations "can help by forcing the researcher to articulate their guiding assumptions, decomposing their study system into the parts, properties, and relationships," but we are not told what a mechanistic explanation actually *is*, whether overt mathematical content, biological plausibility, or some aspect of precision.

Even if it's hard to give a principled definition of *theory* or *mechanism*, perhaps day-to-day usage is sufficiently consistent that it doesn't really matter. I have argued elsewhere that efforts to unpack such terms almost invariably reveal them to depend heavily on authors' particular intellectual and esthetic preferences – which, unsurprisingly, tend to differ widely across individuals (Yarkoni, 2020). We can observe this phenomenon in the present commentaries. Consider the pieces by **Dickins and Rahman** and **Donkin, Szollosi, and Bramley (Donkin et al.)**. Both argue that psychology needs better theory, yet their concrete prescriptions diverge in ways that are not obviously reconcilable. Dickins asserts that grounding psychology in deeper theory requires "seeking some unity with biology, through the adoption of highly corroborated theories such as evolutionary theory"; Donkin et al., by contrast, make no appeal at all to biology or evolution, and instead suggest that the "primary explicanda of psychology are people's capacities," and hence, "[p]sychological explanations should not only account for what people did in some experiment, but also for what they could have done." Of course, these are only two particular positions in a literature replete with differing views as to what skill set or body of knowledge is conducive to good theory.

To be clear, I am not suggesting that pro-theory arguments such as these are wrong, but that they are unhelpful. There are innumerably many reasons why any given statistical result might fail to support a particular verbal claim, and there is little reason to suppose that, say, a social psychologist's failure to consider stimulus variability in an IAT task has much in common with an educational psychologist's failure to consider variability in instructor quality in a study of flipped classrooms. It would be pleasant, but probably wishful thinking to believe we could eliminate most, or even many, generalizability-related problems simply by convincing psychologists to think more about evolution or biology or culture and so forth.

R3.2. Generalizability from a construct validity perspective

Two commentaries approach the issues raised in the target article from a traditional psychometric perspective. **Flake, Luong, and Shaw (Flake et al.)** argue that a productive way forward is to emphasize large-scale construct validation – that is, to conduct extensive descriptive research aimed at ensuring that one's measures are actually measuring what they're supposed to be measuring. **King and Wright** echo this suggestion and further point out that the problems the target article describes in terms of

generalizability can be equivalently construed in terms of construct validity – specifically, the assertion that statistical expressions ought to map closely onto verbal/theoretical expressions can be restated as saying that measures should be valid operationalizations of the constructs they are meant to represent.

I am broadly sympathetic to these commentaries. My only (minor) reservation is a practical one: Framing things in terms of construct validity carries a certain amount of psychometric baggage that, in my view, can be counterproductive. Both **Flake et al.** and **King and Wright** view construct validity as something one ought to establish *before* one starts computing inferential statistics, making predictions, and so on. I think this is good advice for researchers with a realist orientation who construe their research as a search for the latent causes of people's behaviors (for discussion, see Yarkoni, 2020). But this is not the only view one can take. I have argued previously that the data-generating processes underlying many psychological phenomena may simply be too complex and messy for traditional psychometric models to have much utility, so that in practice, the most effective way to make progress may be to largely set aside psychometric concerns about (internal) validity and instead focus more on developing predictively useful models, however complex or uninterpretable they may be Rocca and Yarkoni (in press), Yarkoni and Westfall (2017). I won't defend the latter position here, but am simply observing that the framing I adopted in the target article deliberately sought to minimize theoretical commitments and describe the problem in a maximally general way.

R4. Solutions

The fourth and largest group of commentaries focused on describing one or more solutions to the problems identified in the target article. I have organized these into four subgroups. Respectively, they include commentaries that focus on (1) formal methodologies (either the general need for greater formalism, or specific techniques); (2) benefits afforded by big data and associated technical developments; (3) various methodological procedures, several of which expand on suggestions made in the target article; and (4) bird's-eye or "meta" perspectives that focus on how resources and incentives organize researchers' efforts at a communal level.

R.4.1. Formal methods

Several of the commentaries call for an increased role for formal methodologies in psychological science. **Turner and Smaldino**'s call is the most general; the authors argue for increased emphasis on mechanistic explanation and formal modeling throughout psychology. I have already explained why I find the mechanistic part of their appeal unconvincing; on the contrary, I enthusiastically agree with their call for greater adoption of formal/computational methods. The main reason for this is that I think there is greater transfer between computational skills than between substantive bodies of domain knowledge, and thus computational training of any kind may have more leverage in experimental design and evaluation.

Ross echoes **Turner and Smaldino**'s call for more formal methodology in psychology, and argues, in particular, for greater adoption of methods widely used in experimental econometrics – for example, larger-scale (and more expensive) experiments, and Bayesian estimation. I am sympathetic to many of Ross's specific recommendations, which overlap to some degree with those I made in the target article. That said, I don't think Ross's commentary should be read (or is intended) as an injunction against the use of other modeling techniques and strategies. As **Gelman** points out in his commentary, it may be helpful for scientists to think of statistics as a box of heterogeneous tools, where "[d]ifferent models and statistical tests capture different aspects of the data we observe and the underlying structure we are trying to study." A central message of both commentaries is that it is hubristic to suppose that mindless application of statistical significance tests could produce meaningful answers to most of the questions psychologists pose – so, authors should be prepared to develop a broader toolset.

Braver and Braver and **Bonifay** focus on more specific techniques. Braver and Braver echo my call for a greater focus on variance decomposition approaches, and offer valuable design recommendations (e.g., to try and systematically vary at least one purportedly irrelevant factor in every study). They also take issue with what they see as my unwarranted dismissal of conceptual replications, pointing out that it's entirely possible to aggregate conceptual-related experiments that don't share common design elements via meta-analysis. Although I didn't dismiss conceptual replications, I did observe how difficult it is to integrate the results of conceptual replications in a principled way. One can aggregate estimates from *any* set of studies via meta-analysis. The trouble is that meta-analyses of conceptual replications suffer from the same problem as all other meta-analysis applications: Selection biases which cannot help but reify those biases already baked in by selective reporting and construal. By contrast, explicitly varying multiple design factors within a single study (a strategy that Braver & Braver also endorse) makes it more difficult for authors to mislead themselves.

Bonifay focuses on the minimum description length (MDL) principle as a means of reducing overfitting, and hence (indirectly) also generalization errors. The MDL is one in a class of information theoretic techniques that formally attempt to mitigate overfitting by penalizing models for complexity. Bonifay argues that the MDL "offers insights into overfitting and generalizability that are not possible using traditional methods," and this may be true to some degree. At the same time, I think Bonifay's commentary somewhat oversells the benefits of MDL. Quoting from Grünwald (2005), Bonifay suggests that the MDL "automatically and inherently protects against overfitting." This is somewhat misleading: The MDL is an idealization, and prevents overfitting only in principle. In practice, there is no way to deterministically compute the shortest possible description of a dataset, or even verify that a given proposal is optimal. Specific MDL algorithms *are* computable, but necessarily introduce inductive biases, and hence can and do overfit (they are also restricted to certain classes of models). Moreover, it's important to remember that Occam's Razor (which MDL is a formalization of) is only a heuristic, not a law. The MDL principle offers no guarantee that a favored model adequately captures the true data-generating process, but only that it compactly describes the data. As always, there is no free lunch: Specific MDL algorithms will sometimes perform better than other approaches and sometimes worse, but blanket statements to the effect that the MDL principle overcomes standard problems of model comparison seem to me hard to justify.

Bear and Phillips take issue with the target article's advocacy of more expansive mixed-effects models. They argue that the use of random effects is problematic in many common designs, as the inclusion of such terms depends on the assumption that the levels of the random factors are being sampled randomly from well-

defined underlying populations, which is clearly false in most cases (e.g., most researchers don't really sample their stimuli at random from some well-defined space). I think this argument runs afoul of Box's famous aphorism that "all models are wrong, but some are useful." The point of including random effects is to adjust parameter estimates to account for presumed sources of variance in the data. It should go without saying that in cases where researchers are able to write down a deterministic expression that more closely approximates the true data-generating process, they should do so (see also footnote 11 in the target article). But such scenarios are extremely rare in psychology. The vastly more common scenario involves a choice between a model that makes *no* effort to account for obvious sources of variability in the data, and one that makes an effort to do so imperfectly. Therefore, although Bear and Phillips' titular claim that "random effects won't solve the problem of generalizability" is trivially true, this is hardly a reason to forsake random effects, because the conventional alternative is still worse. A charitable reading of Bear and Phillips is that they are simply pointing out that there can be more suitable formalisms than mixed-effects models in many cases – a view I agree with, and explicitly endorsed in the target article (e.g., "Of course, inclusion of additional random effects is only one of many potential avenues for sensible model expansion").

Finally, **Maniadis** (and to some degree also **Gelman)** provides an important counterpoint to the other commentaries in this group by observing that increased formalism alone will not suffice to solve the generalizability crisis. Maniadis points out that there are other fields (e.g., experimental economics) that already emphasize formal methods to a far greater extent than psychology, yet suffer from very similar problems. This is a point worth reaffirming: Although there can be little doubt that greater statistical sophistication in psychology would improve the state of affairs, it is clearly neither necessary nor sufficient to ensure that researchers produce defensible scientific inferences. Maniadis puts it well in emphasizing the need for caution, observing that "while formalism makes excessive ad hoc theorising more difficult, it does not rule it out."

R.4.2. Benefits of larger, richer datasets

Several commentaries highlight the utility of large, rich datasets in addressing concerns about generalizability, and emphasize the critical role of technology in facilitating the acquisition or analysis of such datasets. I, enthusiastically, endorse the approaches promoted in these commentaries, and have made similar arguments myself in the past (e.g., Yarkoni, 2012a, 2012b). Three of the commentaries focus on the utility of closely related crowdsourcing (**Cyrus-Lai, Tierney, Schweinsberg, & Uhlmann**), "citizen science" (**Hilton & Mehr**), and "many labs" (**Visser et al.**) approaches. The key point here is that establishing the generality of an effect usually requires datasets that sample from a broad universe of observations, and acquiring such datasets is far easier when researchers leverage the scale of internet-based data collection, or join forces and form research consortia. The common goal is to maximize variation in the data – whether by randomly assigning large samples to diverse conditions; by allowing investigators to operationalize hypotheses as they see fit; or by acquiring data at multiple sites, from multiple populations, using multiple methods.

Davidson et al. illustrate how modern technologies – in particular, digital traces of behavior obtained from smartphone sensors and interactions with mobile applications – can be used to expand the scope of measurement of behavior beyond the traditional emphasis on self-report. Although there is much to like about Davidson et al.'s advocacy for the study of digital traces and large-scale data sharing – although, for reasons already alluded to above (see sect. R3.2), their assertion that "it is critically important psychology shifts away from predictive validity alone as evidence for successful operationalization and parameterization" is less attractive.

Finally, **Van de Velde, De Pascale, and Speelman (Van de Velde et al.)** discuss the strengths and limitations of corpus linguistics approaches – which emphasize large, naturalistic datasets over small factorial experiments – when used in pursuit of stronger, more generalizable inferences. Van de Velde et al.'s commentary is notable and refreshing in that the authors emphasize the complexities and tradeoffs involved in adopting corpus linguistics methods, and caution against treating such approaches as a panacea. The point is well taken, and applies well beyond the study of language. The target article provided only a brief sketch of a few modeling strategies that can help close the gap between authors' generalization intentions and their statistical operationalizations; it goes without saying that the central lesson is not that linear mixed-effect models can solve all problems, but rather, that ritualistic reliance on statistical defaults (e.g., the conventional subject-as-random-effect model) rarely leads to good outcomes. Once researchers acknowledge this point, then the difficult work of selecting and specifying a model appropriate to the specific domain and problem at hand can begin – and Van de Velde et al.'s commentary provides a nice case study of the kinds of considerations that may arise.

R.4.3. Methodological recommendations

A third group of solution-focused commentaries consists of what I'll call "methodological recommendations," reflecting their emphasis on a particular type of methodological practice (generally, non-statistical in nature, in contrast to the first subgroup of commentaries in this section).

West, Soska, Cole, Han, Hoch, Hospodar, and Kaplan (West et al.) focus on the role of strictly descriptive work in psychology – that is, research that makes no claim to establish causal relationships, but simply seeks to characterize the relationships between various measured variables. The authors describe several guiding principles that can help improve the quality of descriptive research. I broadly agree with their recommendations. My only minor quibble is that West et al. encourage researchers to thoroughly explore their data before performing inferential tests. Although data exploration is certainly desirable, conditioning one's choice of inferential procedures on prior examination of one's data is an excellent way to procedural overfit that data (Yarkoni & Westfall, 2017). Researchers who wish to follow West et al.'s advice should take pains to maintain a clear separation between exploration and confirmation (e.g., via use of preregistration, hold-out datasets, etc.).

Blersch, Franchuk, Lucas, Nord, Varsanyi, and Bonnell (Blersch et al.) argue for the use of formal causal frameworks as a means of bridging between qualitative and quantitative analysis in psychology. Their commentary echoes other recent appeals for psychologists to embrace causal analysis (e.g., Grosz, Rohrer, & Thoemmes, 2020; Rohrer, 2018). I have mixed feelings about such calls. On the one hand, I agree with the present authors that greater familiarity with the dominant causal approaches

(e.g., Rubin's potential outcomes framework and Pearl's work on causal graphs) might help many psychologists better understand the limitations of their models. On the other hand, I think Blersch et al. considerably overestimate the power of formalisms such as directed acyclic graphs (DAGs) to, as they put it, "bridge between qualitative and quantitative research." The toy example they present in Figure 1 fails to convey what is actually difficult about formalizing causal relationships in most areas of psychology: It isn't the ability to express one's qualitative hypotheses in terms of nodes or edges (witness the rise of closely related structural equation models in psychology over the past few decades), but rather, the ability to justify the assumption that *this* particular graph adequately represents the causal phenomena it is intended to stand in for, in the face of innumerable viable alternatives. Contrary to Blersch et al., such assumptions are usually impossible to test empirically.

Syed and McLean echo the target article's call for more serious consideration of qualitative approaches. A key point the authors emphasize is that a huge amount of the work psychologists currently engage in is already qualitative in nature. Discussion sections unpack the qualitative implications of quantitative results; measurement studies assign qualitative interpretations to factors that are, at bottom, mathematical abstractions; and much of the coded data that enter statistical analyses reflects qualitative assessments. As Syed and McLean observe, "[i]t appears that even qualitative analysis is permissible in mainstream psychology so long as we do not call too much attention to the practice, and do not engage in the intentionality and rigor of best practices in qualitative methods." I strongly endorse Syed and McLean's argument that "qualitative methods can also play a key role in testing, applying, and exemplifying theoretical claims."

Wilford, Ardila-Cifuentes, Baggs, and Anderson (Wilford et al.) argue that concerns about generalizability largely stem from the dominance of the stimulus–response (S–R) paradigm within psychology; they advocate for a different paradigm – the *perturbation experiment* – that avoids these issues. It wasn't clear to me what features the authors think define perturbation experiments, or how such experiments manage to avoid the need to ensure an alignment between one's verbal and statistical statements. On one reading, Wilford et al. are reiterating Popper and Meehl's call for "risky" predictions – for example, they write that perturbation experiments "aim to identify the precise variable or variables implicated in the ongoing control of a complete activity." Accomplishing such a feat would presumably require an experiment to be so carefully operationalized that the outcome rules dispositively in favor of one particular interpretation of a phenomenon. If this is the intended conclusion, I am supportive (and argue as much in the target article). But perturbation as Wilford et al. discuss it doesn't seem either necessary or sufficient for producing risky predictions (e.g., some of the methods Wilford et al. list as intrinsically perturbative in nature, such as TMS and lesion studies, have not precluded problematic conclusions). Moreover, even in the best-case scenario, there remains no escape from the need to align statistical and substantive expressions. To see this, one need to only consider the statistics reported in the elegant Adolph, Eppler, and Gibson (1993) study Wilford et al. hold up as an example of a successful perturbation experiment. Would Wilford et al. continue to argue that the Adolph et al.'s study provides "unambiguous evidence" for its conclusion if it were later discovered that the statistical model had been incorrectly specified, or if the effect were shown to obtain only in the hands of one particular experimenter? It seems doubtful.

R.4.4. Bird's eye views

The last set of solutions-focused commentaries take a bird's eye view of the issues discussed in the target article. Instead of focusing on the mechanics of specific solutions, these commentaries focus on broader cultural issues and incentives, historical perspectives, and cross-field comparisons. Two of the commentaries – **Schiavone, Bottesini, and Vazire (Schiavone et al.)**, and **Sievers and DeFilippis** – argue that the problems described in the target article would be more effectively addressed by focusing on community-level practices and incentives rather than on individual researchers' behavior. I take no position on this claim; the prescriptions I outlined were largely agnostic with respect to implementation. I do, however, think we should be generally wary of arguments to the effect that major cultural changes would, as Schiavone et al. write, "follow swiftly if a small group of gatekeepers decided to make it a priority." It is true that power is disproportionately concentrated in the hands of a relatively few gatekeepers; but gatekeepers generally do not attain their status by operating outside of mainstream mores. Similarly, Siever and DeFilippis's suggestion that we should "foster a diverse scholarly community that is incentivized to reveal what those who came before them have missed" sounds laudable, but implausible in execution.

The **Gigerenzer** and **Alzahawi and Monin** commentaries provide historically oriented perspectives on the field, and travel different paths to arrive at a similar (at least superficially) conclusion: We should think more carefully about how we conduct quantitative research in psychology. Gigerenzer observes that many of our current statistical conventions (e.g., modeling subjects but not stimuli as random effects) reflect historical accidents and misunderstanding of statistics, and suggests we "liberate research practice from methodological rituals." I am very much in agreement with this conclusion, although Gigerenzer's observation that logical fallacies and misunderstandings like the "replication delusion" are widespread even among psychology professors and statisticians raises some serious concerns about the scope of the task at hand.

Alzahawi and Monin's conclusion is, on the surface, similar to **Gigerenzer**'s: The authors suggest that we should work to highlight "how inferential statistics can be more thoughtfully applied." Although the general conclusion is again easy to agree with, the argument Alzahawi and Monin offers in its support is, in my view, self-defeating and rather cynical. The authors specifically reject any effort to move away from quantitative methods in psychology, arguing that such a thing is "unlikely to obtain," because quantitative methods are presently "core to psychology's social and scientific status." This position conflates explanation and justification. It may be true that psychologists historically rushed to adopt quantitative methods in part because doing so conferred prestige and resources on the field; but surely we should not accept it as axiomatic that misaligned incentives cannot ever change, or reform of almost any kind would become impossible. Ironically, Alzahawi and Monin's closing recommendation to "draw more accurate – if more modest – conclusions from our data" is susceptible to their very own argument. There are presently few cultural incentives for psychologists to be more modest in their conclusions or more thoughtful in their inferences; by Alzahawi and Monin's reasoning, shouldn't this doom their own prescription to failure?

Ioannidis takes up the question of how to optimally sequence research activities; specifically, he asks whether it is better to focus on replication first and generalization second, or to do the converse. Ioannidis ascribes to me the latter view – that is, he takes

me to favor a sequence that goes "discover-generalize-replicate, i.e., don't waste time with replication unless a promising research finding has been probed in a sufficiently large variety of settings to have some sense that it is generalizable." He, then, argues that this strategy has downsides, and that there are many scenarios in which it makes sense to try to replicate narrow findings ahead of any attempt to demonstrate their broader generalizability. I agree with this. In writing that "the current focus on reproducibility and replicability risks distracting us from more important, and logically antecedent, concerns about generalizability," I was not suggesting that establishing generality is a more important *empirical* goal than replicability, only that the decision to replicate a given finding should presuppose an adequate understanding of its plausible implications. Researchers who believe it is more important to directly replicate Experiment 1 of Schooler and Engstler-Schooler (1990) than to expand the scope of its design are welcome to privilege the former. But that decision should be made with full, explicit recognition of the experiment's limitations, rather than implicitly or explicitly equating a very narrow operationalization with the broad construct of interest.

Finally, **Lampinen, Chan, Santoro, and Hill (Lampinen et al.)** compare and contrast publishing norms in psychology with those in the field of artificial intelligence (AI), and suggest that each field would benefit from adopting some of the habits of the other. I lack the expertise necessary to evaluate this recommendation with respect to AI, but I largely agree with Lampinen et al.'s suggestion that psychology would benefit from increased adoption of informal, rapid publication streams (cf. Yarkoni, 2012a, 2012b). Of course, it's hard to know how much of AI's rapid progress can be attributed specifically to its publishing conventions; I have previously suggested that much of machine learning and AI's success may stem from its emphasis on evaluating models against standardized benchmarks – a practice that contrasts markedly with psychologists' tendency to choose their own idiosyncratic evaluation metrics on a case-by-case basis (Rocca & Yarkoni, in press). But, either way, I agree with Lampinen et al.'s concrete recommendations.

Notes

1. The "Author's Response" was edited, with my permission, for tone and content by Dr. Barbara Finlay.
2. For context, the concerns raised here Lakens, Uygun Tunç, and Tunç are like those brought up by Lakens in a much longer open commentary (http://daniellakens.blogspot.com/2020/01/review-of-generalizability-crisis-by.html) on an earlier draft of the target article, to which I wrote a detailed online rebuttal (https://www.talyarkoni.org/blog/2020/05/06/induction-is-not-optional-ifyoure-using-inferential-statistics-reply-to-lakens/).

References

Adolph, K. E., Eppler, M. A., & Gibson, E. J. (1993). Crawling versus walking infants' perception of affordances for locomotion over sloping surfaces. *Child Development 64*(4), 1158–1174.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J.…, Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy 22*(4), 421–435.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science 15*(5), 1243–1255.

Grünwald, P. (2005). A tutorial introduction to the minimum description length principle. In P. Grünwald, I.J. Myung, and M. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 3–81). MIT Press.

Rocca, R., & Yarkoni, T. (in press). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *PsyArXiv*. https://doi.org/10.31234/osf.io/e437b.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science 1*(1), 27–42.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology 22*(1), 36–71.

Yarkoni, T. (2012a). Designing next-generation platforms for evaluating scientific output: What scientists can learn from the social web. *Frontiers in Computational Neuroscience 6*(October), 72.

Yarkoni, T. (2012b). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science 21*(6), 391–397.

Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on Fried (2020). *Psychological Inquiry 31*(4):326–333.

Yarkoni, T., & Westfall, J. (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science 12*(6):1100–1122.